Statistics Canada Statistique Canada

SURVEY METHODOLOGY

TECHNIQUES D'ENQUÊTE

June - 1977 - Juin

VOLUME 3

NUMBER 1 - NUMÉRO 1

A Journal produced by
Statistical Services Field
Statistics Canada

Publié par Le Secteur des
Services Statistiques,
Statistique Canada

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1977 Vol. 3 No. 1

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

C O N T E N T S

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1977 Vol. 3 No. 1

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Editorial Policy:

The objective of the Survey Methodology Journal is to provide a forum in a Canadian context for publication of articles on the practical applications of the many aspects of survey methodology. The Survey Methodology Journal will publish articles dealing with all phases of methodological development in surveys, such as, design problems in the context of practical constraints, data collection techniques and their effect on survey results, non-sampling errors, sampling systems development and application, statistical analysis, interpretation, evaluation and interrelationships among all of these survey phases. The emphasis will be on the development strategy and evaluation of specific survey methodologies as applied to actual surveys. All papers will be refereed, however, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or the Department. Copies of papers in either Official Language will be made available upon request.

Politique de la rédaction:

La revue Techniques d'enquête veut donner aux personnes qu'intéressent les aspects pratiques de la conduite d'enquetes, la possibilité de publier sur ce sujet dans un cadre canadien. Les textes pourront porter sur toutes les phases de l'élaboration des méthodes d'enquête: les problèmes de conception causé par des restrictions pratiques, les techniques de collecte de données et leur incidence sur les résultats, les erreurs d'observation, l'élaboration et l'application de systèmes d'échantillonnage, l'analyse statistique, l'interprétation, l'évaluation et les liens entre les différentes phases d'une enquête. On s'attachera principalement aux techniques d'élaboration et à l'évaluation de certaines méthodologies appliquées aux enquêtes existantes. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne seront pas nécessairement celles du comité de rédaction ni de Statistique Canada. On pourra se procurer sur demande des exemplaires d'un article dans l'une ou l'autre langue officielle.

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

June/juin 1977 Vol. 3 No. 1

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

Submission of Papers:

The Journal will be issued twice a year. Authors are invited to submit their papers, in either of the two Official Languages, to the Editor, Dr. M.P. Singh, Household Surveys Development Division, Statistics Canada, 10th Floor, Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Two copies of each paper, typed space-and-a-half, are requested. Authors of articles for this journal are free to have their articles published in other statistical journals.

Présentation de documents pour publication:

La revue sera publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à en faire parvenir le texte au rédacteur en chef, M. M.P. Singh, Division de l'élaboration d'enquêtes ménages, Statistique Canada, 10e étage, Edifice Coats, Tunney's Pasture, Ottawa, Ontario, K1A 0T6. Prière d'envoyer deux exemplaires, dactylographiés à interligne et demi. Les auteurs des articles publiés dans cette revue sont libres de les faire paraître dans d'autres revues statistiques.

ADJUSTMENT FOR NON-RESPONSE IN SURVEYS

R. Platek, M.P. Singh and V. Tremblay
Household Surveys Development Division

The paper attempts to examine some of the procedures used for compensation for non-response. Using the concept of response probabilities, a simple response - non-response error model is developed and the components of response and non-response errors are identified under various imputation procedures. A graph is also given in order to provide an idea of the magnitude of the non-response bias in a particular situation. Two examples of the practical application of imputation are discussed.

1. INTRODUCTION

Any survey, whether it is a sample survey or a census, suffers from some non-response. Most practicing statisticians or data analysts will agree that disregarding non-response may lead to survey results of poor quality and should only be considered as a last resort. The proper method of dealing with non-response depends upon the type of survey and the size of non-response. In analytical surveys, for example, the important issue may not be the potential bias in the estimate of means, medians or percentages but rather the potential bias in relationships between the variables under analysis. Thus, the method of dealing with non-response depends upon the degree of interaction between the variables in the relationship and the rate of non-response, but not purely upon the non-response.

Most surveys, however, produce basic population estimates, which are usually in several dimensions and which pertain to various types of geographic areas. In these surveys, the elimination or reduction of non-response is very important. The size of the non-response depends both upon the subject matter and the method of data collection. Some questions which happen to be of a more personal nature like income, health, etc, generate more non-response than other less sensitive questions. The use

of interviewers in face to face interviews generally produces less non-response than the use of telephone or mail surveys but at a greater cost per interview. In addition, non-respondents may be of various types - not at home, refusal, etc. At the collection stage, the 'not at home' rate can be reduced by persistent efforts of interviewers. The refusal rate may be reduced by changing interviewers, eg. a different interviewer may succeed where the first failed, or by introducing special publicity campaigns for the survey. But however successful the field operations may be, there is a point beyond which non-response cannot be further reduced at reasonable cost. Just as a substantial increase in the sample size is required to effect even a marginal reduction in the mean square error of statistics after a certain point, so enormous cost and effort would be needed to remove the last few traces of non-response. Thus, in practically all surveys some form of adjustment for non-response must be applied and in such a manner as to take into account the differences in the characteristics of respondents and non-respondents by utilizing relevant information available about the two groups.

In this paper, an attempt is made to summarize some of the procedures commonly used for compensation of non-response and to identify and isolate the non-response error components according to the methods used for non-response adjustments. A graph is also provided which may be used to obtain an idea about the magnitude of non-response bias in a given survey. Two Canadian studies using different imputation procedures are outlined indicating anticipated non-response bias by characteristics.

2. METHODS OF COMPENSATION FOR NON-RESPONSE

We briefly outline below some methods of compensation commonly used. Detailed descriptions of the methods are given in Chapman (1976).

A. No adjustment: In such cases data presented refer only to the responding group and no compensation is made for non-respondents.

B. Design dependent balancing factor: This is the simplest method of non-response adjustment. This adjustment may be applied at any given sample design level, such as, primary sampling unit, stratum,

group of strata or entire sample. Such an adjustment is achieved by application of a balancing factor to the design weight at the given design (balancing) level. Normally, this factor is simply the ratio of sample size to the number of respondents at the balancing level. In addition to being the simplest procedure, if the characteristics of respondents and non-respondents are similar at the balancing level, then this procedure would be preferable in the sense of relative bias as well. The choice of balancing level is, however, quite critical in this method as the bias would differ for different choices of balancing levels.

C. Creation of post-strata (or weighting classes): This procedure is based on the principle of post-stratification (stratification after sampling), where strata are formed on the basis of information obtained from the characteristics from the sample. A variation of the method of forming the weighting classes is discussed by Morgan and Sonquist (1963). A uniform non-response weight is applied within each class. Again, the choice of characteristics and the size of classes are quite important as the bias would depend upon the homogeneity of characteristics between the respondents and non-respondents, and the non-response rate within the classes.

D. Similar record substitution: After forming the weighting class or balancing units as in C or B above, each missing record is replaced by a record selected at random (or following some systematic rule) from the set of respondents within the same class.

E. Similar respondent substitution: For some surveys, non-responding units are replaced in the field by selecting substitute units from the population. An attempt is made in this procedure to substitute the non-responding units by units having similar characteristics.

F. Use of external data for non-respondents: Adjusting for non-response may be based on the information obtained from the census or earlier surveys or in the case of rotating samples, from previous interviews. This method could be combined with method B or C when historical data are available only for some rather than all non-respondents.

G. Weighting or substitution on the basis of follow-up results: Whenever one or more follow-ups on non-respondents is exercised, the analysis of the characteristics of the respondents in the first, second, ..., occasion permits, to some extent, inferring on the characteristics of hard core non-respondents.

H. Adjustment of the macro-data: Sometime, it may not be possible to post-stratify as suggested in C, because the auxiliary information is not available at the individual non-responding unit level but rather at an aggregate level. In such cases, adjustment factors may be calculated to make the estimate consistent with the external source. If data are available from more than one source, then the adjustment may be performed through the application of raking ratio procedures.

3. FORMULATION OF THE NON-RESPONSE BIAS

For the purpose of this discussion, we shall restrict ourselves to the complete enumeration situation and we consider estimation of the population total X for some given characteristic "x". The entire population may be classified into "respondents" and "non-respondents" as in the following table.

	Respondents	Non-Respondents	Total Population
Total for "x"	X_R	X_N	$X=X_R+X_N$
Number of Units	R	N	T=R+N

Here, X_N (hence X) is unknown but must be imputed using one of the methods previously described. Let us denote by Z_N the value imputed for X_N; then $\hat{X}=X_R+Z_N$ will be the estimate used for X. In general, the bias of X may be expressed as

$$B(\hat{X}) = \hat{X} - X = Z_N - X_N = N\cdot(\bar{Z}_N - \bar{X}_N) \quad (3.1)$$

where $\bar{X}_N$ is the average for the non-responding groups.

In the case of procedure A, where no compensation is made for non-respondents, the bias is the same as the total (X_N) for the non-respondents, that is, the coverage of the survey in such cases gets restricted to the responding group only.

Now, if a uniform correction factor is applied at the entire population level, then the average $\bar{X}_R$ for the respondents is imputed for non-respondents and

$$B(\hat{X}) = N\cdot(\bar{X}_R - \bar{X}_N) \ . \tag{3.2}$$

If the correction factor is applied uniformly within each stratum or post-stratum "h", then the bias expression becomes

$$B(\hat{X}) = \sum_h N^{(h)} \ (\bar{X}_R^{(h)} - \bar{X}_N^{(h)}). \tag{3.3}$$

This implies that estimate (3.3) will be less biased than the estimates given by (3.2) if respondents and non-respondents are more alike within strata or post-strata than in the whole population.

Further, if the missing data are imputed using external data on the non-respondents providing an overall mean $\bar{X}_N'$ as a proxy for $\bar{X}_N$, then the bias may be expressed as a simple function of the errors in the external data for the non-responding units, namely

$$B(\hat{X}) = N(\bar{X}_N' - \bar{X}_N). \tag{3.4}$$

In many practical situations, the survey statistician is not in a position to estimate the non-response biases. However, it would be highly desirable for data analysis purposes to establish an upper bound (positive or negative) for the biases. For example, whatever the method of correction for non-response, it is likely that $\bar{Z}_N$ will be at least as close or closer to $\bar{X}_N$ than is $\bar{X}_R$; that is, in the mind of the statistician the inequality

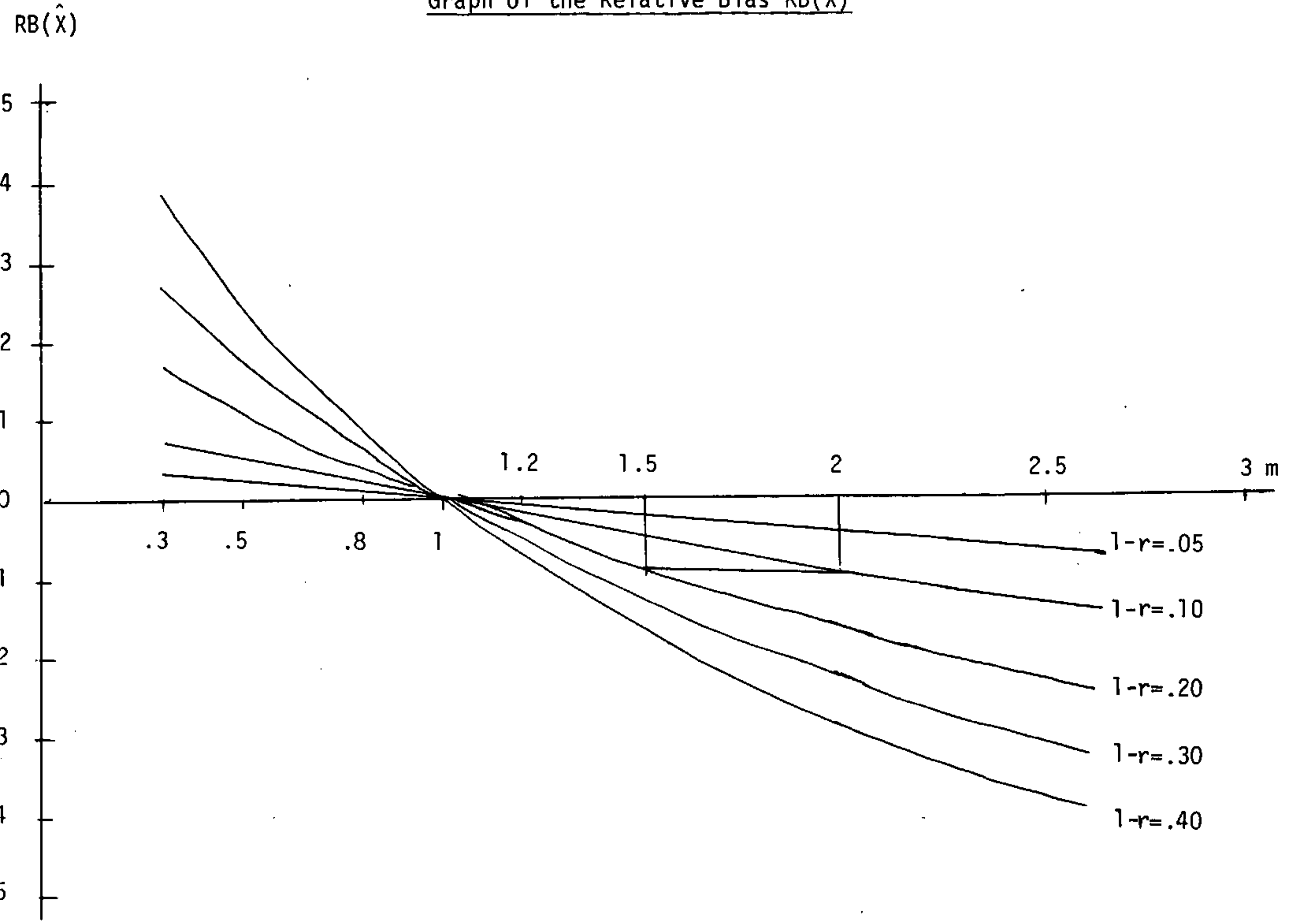
$$| \ \bar{Z}_N - \bar{X}_N \ | \leq | \ \bar{X}_R - \bar{X}_N \ | \tag{3.5}$$

is always satisfied. If that was not the case, one would apply the least expensive and the simplest method of compensation for non-response (by estimating $\bar{X}_N$ by $\bar{X}_R$). In that case, an upper limit of bias may be obtained by using $B(\hat{X})$ in (3.2) and the relative bias of $\hat{X}$ in this case may be expressed as

$$RB(\hat{X}) = \frac{\hat{X}-X}{X} = \frac{(1-r)(1-m)}{r+m(1-r)} = \frac{(1-r)(1-m)}{1-(1-r)(1-m)} \qquad (3.6)$$

where $m = \bar{X}_N/\bar{X}_R$ and $1-r$ is the non-response rate. For a given value of $1-r$, the relative bias $RB(\hat{X})$ is simply a function of m and it is shown graphically how the relative bias is affected by changes in m.

Graph of the Relative Bias $RB(\hat{X})$



Thus, if one has some idea about the value of m based on earlier censuses or surveys or from follow-up studies, then it would be possible to obtain the magnitude of relative bias with the help of the above graph. In other words, if some information on m is available, then one can determine the target response rate to be achieved in a given survey so that the bias is within tolerable limits. It should, however, be noted that if the procedures for reduction of the non-response rate are not well thought out and inappropriately executed aimed at reducing the number of non-respondents, then it may in some circumstances have an adverse effect. For example, if under normal survey conditions $1-r = .20$ and if $m = 1.5$, then $RB(\hat{X}) = -.09$. If by some special inappropriate procedure $(1-r)$ was reduced to .10 but m increased to 2, then $RB(\hat{X}) = -.09$ and the added expenses of reducing non-response results in no reduction in the non-response bias. Further, if $m > 2$, $RB(\hat{X})$ would have become larger than -.09.

4. AN ALTERNATIVE APPROACH TO THE NON-RESPONSE PROBLEM

4.1 The Concept of Response Probability

It has been emphasized above that the problems due to non-response are caused by the differences in characteristics between non-respondents and respondents. Consequently, in order to overcome these problems, survey statisticians attempt to obtain some indications of the characteristics of the non-respondents by such means as follow-up, record substitution, and linkage with other sources.

In this section, we propose to approach the problem using the concept of response probability. It is based on the natural assumption that in the population, each unit has a given probability of responding (if selected) given the survey conditions. In most of the situations, the population is not formed of respondents and non-respondents, but rather of a collection

of potential respondents some of which are respondents and the others non-respondents depending on the conditions under which the survey is conducted. This concept is implicit in the application of the Politz-Simmons Technique (1949). Let us first examine the possible correspondence between the approach based on the differences in characteristics and the approach which takes into account differences in response probabilities for a particular characteristic. Consider a simple situation of complete enumeration of a dichotomous population of size T for the purpose of estimating the total number of units (X) having a particular characteristic "x". If, at the time of enumeration, N non-responses were encountered, the entire population may be then described as follows.

	Respondents	Non-Respondents	Total
Having the Characteristic	X_R	X_N	X
Not Having the Characteristic	$\tilde{X}_R$	$\tilde{X}_N$	$\tilde{X}$
Total	R	N	T=R+N

The parameter X may be expressed in the following two ways:

(a) $(X/T)T$ or (b) $(X/X_R)X_R$

Using (a), the proportion X/T could be estimated by (X_R/R) under the assumption that the respondents and non-respondents have the same characteristics. Alternatively while using (b), the ratio (X/X_R) which is the inverse of the response rate for those having the characteristics, may be estimated by the overall inverse response rate given by T/R. The assumption in case (b) being that the response rates (or response probabilities) are the same for the two categories, namely, those 'having the characteristic' or 'not having the characteristic'. The estimate of the parameter X could be written as $\hat{X} = (X_R/R).T$ under (a) which is obviously the same as that arrived under (b) namely $(T/R)X_R$, establishing

the equivalence between the two approaches. It may be noted that such equivalence exists under other methods of correction for non-response, implying that any assumption made on (a) would lead to a corresponding assumption under (b).

If the assumptions made for estimating X do not hold, the estimate is biased and the non-response bias corresponding to (a) and (b) may be expressed as

$$B_{(a)} = N\left(\frac{X_R}{R} - \frac{X_N}{N}\right) \tag{4.1}$$

and

$$B_{(b)} = \frac{X\tilde{X}}{R}\left(\frac{X_R}{X} - \frac{\tilde{X}_R}{\tilde{X}}\right) \tag{4.2}$$

respectively. It can be shown that $B_{(a)} = B_{(b)}$.

Comparing these two statements, it is clear that the non-response bias is, following (a), directly proportional to the differences between the proportions of respondents and non-respondents having the characteristic or, following (b), it is directly proportional to the differences of response rates between the units having the characteristic and the units which do not. Further, looking at $B_{(a)}$ and $B_{(b)}$, it can be seen that the condition for the non-response bias to be positive may be written either as $X_R/R > X_N/N$ or as $X_R/X > \tilde{X}_R/\tilde{X}$. This can be interpreted as "more respondents would have the characteristic than non-respondents" or "units having the characteristic are more inclined to respond".

4.2 A Simple Response - Non-Response Error Model

Using the concept of response probability, we have developed expressions for response and non-response errors for various imputation procedures. We shall restrict ourselves to a complete enumeration case and consider estimation of the population total

$$X = \sum_{i=1}^{T} x_i \quad .$$

The following two step procedure is applied for each unit in the target population.

Step 1: to identify and contact the respondent, and to seek his co-operation;

Step 2: to obtain a correct response from the respondent, or, if no respondent is available, to find an appropriate substitute for the missing response.

For a particular unit U_i, let x_i denote the true value, y_i the observed value, if the unit is responding, and z_i the imputed value (using one or the other procedure of imputation), if the unit is not responding. Making use of the random variable δ_i which takes the value "1" if the unit is responding and "0" otherwise, the estimate for x_i may be written as

$$\hat{x}_i = \delta_i \; y_i + (1-\delta_i) \; z_i \tag{4.3}$$

where, of course, y_i and z_i are subject to errors. Thus, the response non-response error model may be expressed as

$$\hat{x}_i = \delta_i \; (x_i + {}^{R}\varepsilon_i) + (1-\delta_i)(x_i + {}^{NR}\varepsilon_i) \tag{4.4}$$

where ${}^{R}\varepsilon_i$ and ${}^{NR}\varepsilon_i$ are respectively random errors due to response and non-response. The distribution of δ_i and the random errors are as follows:

$$E_1 \; \delta_i = p_i, \text{ the response probability for unit } U_i$$

$$Var_1 \; \delta_i = p_i \; (1-p_i), \; Cov_1(\delta_i, \delta_j) = 0, \; i \neq j,$$

where E_1 and Var_1 are expectation and variance for step 1 of the procedure mentioned above. Further, denoting E_2 and Var_2 as the expectation and variance for step 2 we have

$$E_2 \; {}^{R}\varepsilon_i = {}^{R}B_i, \quad \text{the response bias for unit } U_i$$

$$Var_2 \; {}^{R}\varepsilon_i = {}^{R}\sigma_i^2, \quad \text{the simple response variance for unit } U_i$$

with the assumption that $\mathrm{Cov}_2\ {}^{R}\varepsilon_i,\ {}^{R}\varepsilon_j = 0$ when $i \neq j$. The assumptions that $\mathrm{Cov}_1(\delta_i, \delta_j)$ and $\mathrm{Cov}_2\ {}^{R}\varepsilon_i,\ {}^{R}\varepsilon_j$ are both zero have been made here to simplify the algebra. In practice, the covariances may not be zero, if for example, the sampling units are observed by the same interviewer.

Further, $E_2\ {}^{NR}\varepsilon_i = {}^{NR}B_i$ imputation bias for unit U_i

$\mathrm{Var}_2\ {}^{NR}\varepsilon_i = {}^{NR}\sigma_i^2$ imputation variance for unit U_i.

The total $X = \sum_i x_i$ is estimated by $\hat{X} = \sum_i \hat{x}_i = \sum_i \delta_i\ y_i + (1-\delta_i) z_i$,

where the z_i's take a particular meaning under each method of compensation for non-response.

We now develop general expressions for the bias and variance of $\hat{X}$ under any imputation procedures when considering the simple response non-response error model described above. Writing $\hat{x}_i = x_i + \delta_i\ {}^{R}\varepsilon_i + (1-\delta_i)\ {}^{NR}\varepsilon_i$, the bias $B(\hat{X})$ is given by

$$B(\hat{X}) = E\ \hat{X} - X = \sum_i E(\hat{x}_i - x_i) = \sum_i E[(1-\delta_i)\ {}^{NR}\varepsilon_i + \delta_i\ {}^{R}\varepsilon_i]$$

$$= \sum_i E_1(1-\delta_i)\ E_2\ {}^{NR}\varepsilon_i + E_1\ \delta_i\ E_2\ {}^{R}\varepsilon_i$$

$$= \sum_i E_1(1-\delta_i)\ {}^{NR}B_i + \sum_i p_i\ {}^{R}B_i \qquad (4.5)$$

$$= \text{Non-Response Bias} + \text{Response Bias}.$$

The general expression for the variance of $\hat{X}$ is

$$V(\hat{X}) = \mathrm{Var} \sum_i [\delta_i\ {}^{R}\varepsilon_i + (1-\delta_i)\ {}^{NR}\varepsilon_i]$$

$$= [\mathrm{Var}_1\ E_2 + E_1\ \mathrm{Var}_2] \sum_i [\delta_i\ {}^{R}\varepsilon_i + (1-\delta_i)\,{}^{NR}\varepsilon_i]$$

$$= \mathrm{Var}_1 \sum_i [\delta_i\ {}^{R}B_i + (1-\delta_i)\ {}^{NR}B_i]$$

$$+ E_1\ \mathrm{Var}_2 \sum_i \delta_i\ {}^{R}\varepsilon_i + E_1\ \mathrm{Var}_2 \sum_i (1-\delta_i)\ {}^{NR}\varepsilon_i$$

$$+ E_1\ \mathrm{Cov}_2 \sum_i \delta_i\ {}^{R}\varepsilon_i, \sum_j (1-\delta_j)\ {}^{NR}\varepsilon_j \qquad (4.6)$$

= non-response variance

+ response variance + imputation variance

+ covariance between the response and imputation errors.

Now, the bias and variance expression under some of the imputation methods described earlier are as follows:

1) If no adjustment is made to correct for non-response, then ${}^{NR}\varepsilon_i = -x_i$ with probability 1 for each non-responding unit U_i and it follows that

$$B_1 = -\sum_i (1-p_i)\, x_i + \sum_i p_i\, {}^{R}B_i \tag{4.7}$$

and

$$V_1 = \mathrm{Var}_1 \sum_i [\delta_i\, {}^{R}B_i - (1-\delta_i)\, x_i]$$

$$+ E_1 \sum_i \delta_i\, {}^{R}\sigma_i^2$$

$$= \sum_i (x_i + {}^{R}B_i)^2\, p_i\, (1-p_i) + \sum_i p_i\, {}^{R}\sigma_i^2 \quad . \tag{4.8}$$

2) If a uniform correction factor (i.e. the inverse of the response rate is applied to each record, then this technique is equivalent to replace each missing record by the observed average $\bar{y}_R$. The error made if unit U_i is not responding is then

$${}^{NR}\varepsilon_i = \sum_i \delta_i\, y_i / \sum_i \delta_i - x_i$$

$$= \bar{y}_R - x_i \quad .$$

It follows that

$${}^{NR}B_i = \sum_i \delta_i\, (x_i + {}^{R}B_i) / \sum_i \delta_i - x_i$$

and $$^{NR}\sigma_i^2 = \mathrm{Var}_2\ ^{NR}\varepsilon_i = \mathrm{Var}_2\ \bar{y}_R = \frac{\sum_i \delta_i\ ^R\sigma_i^2}{(\sum_i \delta_i)^2}$$

neglecting terms of the order of the relative variance of $(\sum_i \delta_i)$, the unbiased estimate of the expected number of responses. In this case then,

$$B_2 = E_1 \sum_i (1-\delta_i)[\sum_i \delta_i(x_i + {}^RB_i)/\sum_i \delta_i - x_i] + \sum_i p_i\ ^RB_i$$

$$= (1-\bar{p})/\bar{p}\ \sum_i p_i(x_i + {}^RB_i) - \sum_i (1-p_i)\ x_i + \sum_i p_i\ ^RB_i$$

$$= [\sum_i x_i\ (p_i/\bar{p}-1) + (1-\bar{p})/\bar{p}\ \sum_i p_i\ ^RB_i] + \sum_i p_i\ ^RB_i \qquad (4.9)$$

where $\bar{p}$ is the expected response rate.

Comparing B_1 with B_2, it is clear that a significant reduction of the non-response bias may be achieved when applying the balancing method when the units have more or less the same probability of responding. The second term in the non-response bias component, (that is $(1-\bar{p})/\bar{p} \sum_i p_i\ ^RB_i$), however, shows, when using this technique, how response errors may contribute to non-response biases. It is as if response errors were inflated by a factor of $\bar{p}^{-1}$; this is to be expected since individual responses are here inflated by a factor of $(T/\sum_{i=1} \delta_i)$.

In this case, the variance (V_2) expressed as a sum of the four components (equation 4.6) is given by

$$V_2 = \mathrm{Var}_1\ \{\sum_i \delta_i\ ^RB_i + (1-\delta_i)[\sum_i \delta_i(x_i + {}^RB_i)/\sum_i \delta_i - x_i]\}$$

$$+ E_1 \sum_i \delta_i\ ^R\sigma_i^2 + E_1\ [\sum_i (1-\delta_i)/\sum_i \delta_i]^2 \sum_i \delta_i\ ^R\sigma_i^2$$

$$+ 2\ E_1\ [\sum_i (1-\delta_i)/\sum_i \delta_i] \sum_i \delta_i\ ^R\sigma_i^2.$$

On simplification, V_2 may be expressed as

$$= \mathrm{Var}_1 \left[\frac{\sum_i \delta_i \, (x_i + {}^R B_i)}{\sum_i \delta_i / T}\right] + E_1 \frac{\sum_i \delta_i \, {}^R\sigma_i^2}{(\sum_i \delta_i / T)^2}$$

which, when using the usual approximation, is equal to

$$V_2 = \bar{p}^{-2} \{\mathrm{Var}_1 \, [\sum_i \delta_i \, (x_i + {}^R B_i) - \bar{y} \sum_i \delta_i] + \sum_i p_i \, {}^R\sigma_i^2\}$$

$$= \bar{p}^{-2} \{\sum_i (x_i + {}^R B_i - \bar{y})^2 \, p_i(1-p_i) + \sum_i p_i \, {}^R\sigma_i^2\} \qquad (4.10)$$

where $\bar{y} = \sum_i p_i(x_i + {}^R B_i)/\sum_i p_i$ is

the average expected observation under the given response - non-response errors conditions. Given the compensation factor $(T/\sum_{i=1}^{N} \delta_i)$, the variance is inflated by the factor of $\bar{p}^{-2}$; but, significant reduction of the non-response component of the variance may be achieved when dealing with homogeneous populations.

3) As a refinement of this method, the compensation factor may be calculated separately for each weighting class "h" (strata or post-strata). The bias and variance expressions are

$$B_3 = \sum_h \bar{p}_h^{-1} \, [\sum_i x_{hi} \, (p_{hi} - \bar{p}_h) + (1 - \bar{p}_h) \sum_i p_{hi} \, {}^R B_{hi}] \qquad (4.11)$$

$$+ \sum_h \sum_i \, p_{hi} \, {}^R B_{hi}$$

and

$$V_3 = \sum_h \bar{p}_h^{-2} \, [\sum_i (x_{hi} + {}^R B_{hi} - \bar{y}_h)^2 \, p_{hi}(1 - p_{hi}) + \sum_i p_{hi} \, {}^R\sigma_{hi}^2] \qquad (4.12)$$

where $\bar{p}_h$ and $\bar{y}_h$ are respectively the expected response rate and the average expected observation within weighting class "h". Then two expressions explain how the criterion of homogeneity within weighting class from both the characteristic and the response probability point of view supports the use of such a method.

4) As a last illustration, we consider the use of external data as a substitute for the missing responses. If for each non-response there exists some auxiliary data available from an external source, which might be also affected by response errors, then the non-response error made for the missing unit U_i is

$$^{NR}\varepsilon_i = z_i - x_i$$

where z_i is the substitute response.

Denoting by z_i^* the expected response over the distribution of possible responses to the external source from unit U_i, the non-response bias $^{NR}B_i (= z_i^* - x_i)$ is comprised of two components: the response bias from the external source and the conceptual differences between the two sources of information. Also, we shall define the non-response variance for unit U_i as $^{NR}\sigma_i^2$, which is simply the response variance of z_i. If the response errors in the external source are not correlated, it would follow that Cov $^{NR}\varepsilon_i$, $^{NR}\varepsilon_j$ = 0 for $i \neq j$. Substituting these values in the general expressions, we find

$$B_4 = \sum_i E_1 \ (1-\delta_i)(z_i^* - x_i) + \sum_i p_i \ {}^{R}B_i$$

$$= \sum_i \ (1-p_i)(z_i^* - x_i) + \sum_i p_i \ {}^{R}B_i \qquad (4.13)$$

and

$$V_4 = \mathrm{Var}_1 \ \sum_i [\delta_i \ {}^{R}B_i + (1-\delta_i)(z_i^* - x_i)]$$

$$+ E_1 \sum_i \delta_i \ {}^{R}\sigma_i^2 + E_1 \sum_i \ (1-\delta_i) \ {}^{NR}\sigma_i^2$$

$$= \sum_i (x_i + {}^R B_i - z_i^*)^2 \, p_i (1-p_i)$$

$$+ \sum_i p_i \, {}^R\sigma_i^2 + \sum_i (1-p_i) \, {}^{NR}\sigma_i^2 \quad . \qquad (4.14)$$

These expressions show how the efficiency of such a technique is a function of the closeness of the two sources of information. As a special case, one may think of using the previous month's data whenever available in a continuing survey (see Section 5.2). The non-response bias for the stable characteristics would be simply a function of the previous month's response bias; if further, response errors tend to be consistent over time, the non-response variance component would become negligible.

Extensions of this approach may be considered by relaxing some of the underlying hypothesis of the error model. Another use of the response - non-response error model is to study how these two error components interact under any imputation procedure when the data collection method is altered. For example, if proxy responses are introducing significant errors, then the procedure might be to accept non-proxy responses only. This would reduce the response error component with the risk of reducing the response probabilities for the units hard to contact. As another example of this interaction, one might think of the application of special procedures to diminish the number of refusals in a sensitive survey; care has to be exercised in such a situation as not to off-set the increase of response probabilities by the introduction of larger response errors. This would occur if reluctant respondents are more prone to give inaccurate or wrong answers.

5. SOME EMPIRICAL RESULTS

We present in this section examples from two surveys, namely (i) Auto Exit Survey and (ii) Canadian Labour Force Survey. The auto exit survey was a one shot pilot survey using mail-back and a follow-up procedure without any control on the data collection mechanism whereas the Labour Force Survey is a regular monthly survey utilizing well organized and trained interviewing staff. The subject matter content, non-response rates, imputation procedures and handling of non-respondents are very much different in the two cases. These two different surveys briefly described below may be considered as two examples with regard to non-response biases and their effect on the estimates derived from these surveys.

5.1 Auto Exit Survey

A pilot survey was conducted at Ontario-U.S. land border points in order to obtain data on travel characteristics (especially length of stay and expenditures) of U.S. visitors leaving Canada by private auto. The sample consisted of three panels. The questionnaires were handed out to the visitors to be mailed back. For one of the panels, addresses of the visitors were obtained and non-respondents were followed up by sending another questionnaire to be mailed back. For this panel 553 questionnaires (blue) were distributed to those persons having spent at least one night in Ontario and 287 responded (52% response) on the first occasion. The remaining 266 were followed up by sending another questionnaire (pink) of which 105 responded (39% response), giving an overall response rate of 71%. The two sets of questionnaires - blue and pink, were separately tabulated for three items: namely, (a) mean nights in Ontario, (b) expenditure per party and (c) expenditure per party night, all data relating to overnight visitors only. Denoting by

$\bar{Y}_1$: average over first occasion respondents

$\bar{Y}_2$: average over second occasion respondents

$\bar{Y}_3$: average over hard core non-respondents,

consider the following general rule for imputing $\bar{Y}_3$:

$$\bar{Y}_3 = (1-a)\ \bar{Y}_1 + a\ \bar{Y}_2 \quad 0 \leq a \leq 2$$

where a is constant, the value of which depends upon the underlying hypothesis about the non-respondents. We consider the following values of a for illustration and the resultant hypotheses are thus as follows.

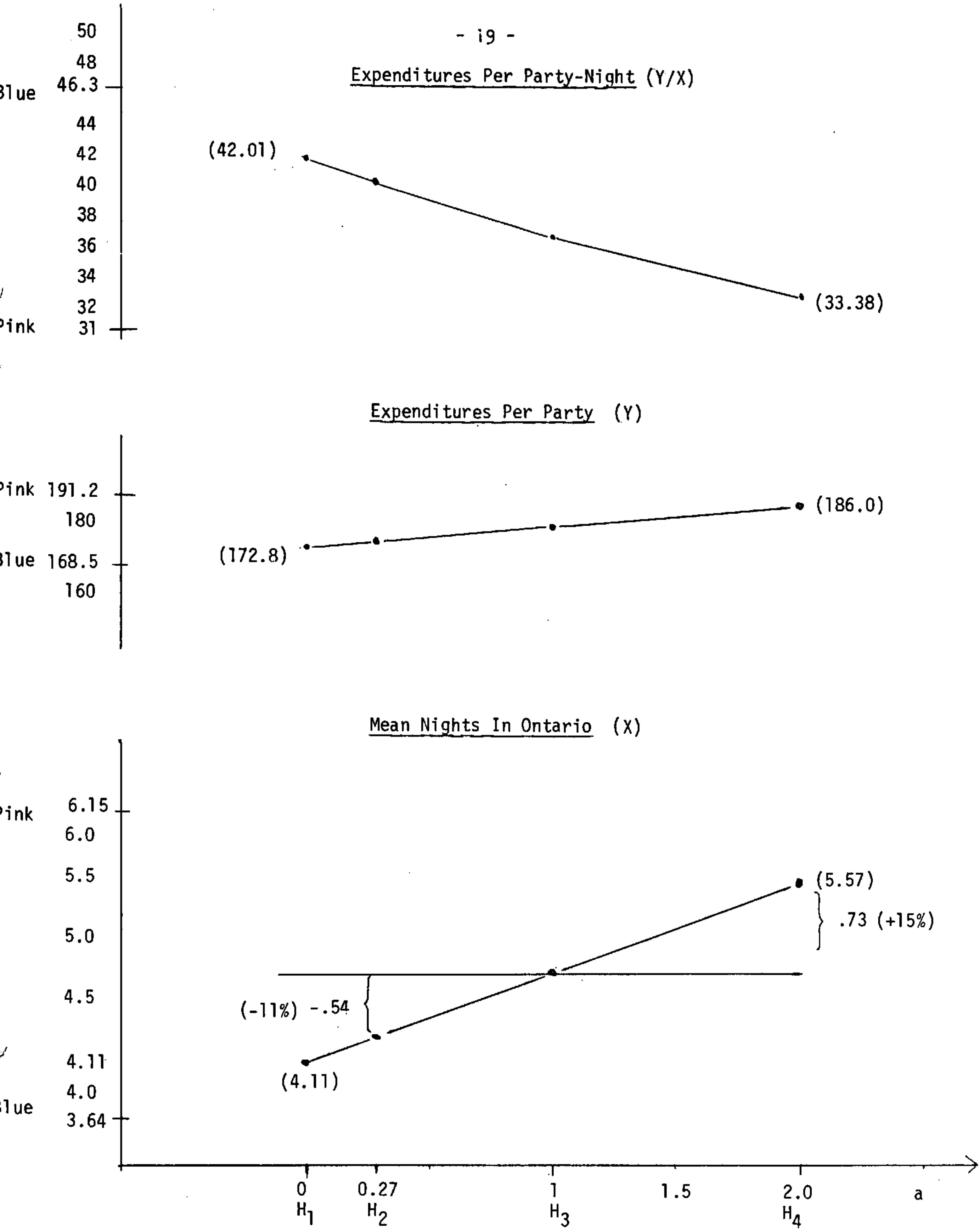
i) $a = 0$: H_1: The first (blue) response represents the hard core non-response.

ii) $a = \frac{n_2}{n}$: H_2: Blue and Pink combined (both response stages) represent the hard core non-response; n_2 is the number of second occasion respondents and n is the total number of respondents.

iii) $a = 1$: H_3: The second (pink) response represents the hard core non-response.

iv) $a = 2$: H_4: The population consists of three 'equally spaced' segments, that is, there is the same difference from pink to non-response as from blue to pink.

The graphs on the following page illustrate the variations in the estimates for three characteristics, the four hypotheses are results of the step-by-step application of method G, based on follow-up responses. These graphs reflect the importance of making a judicious choice of the underlying hypothesis, especially for those characteristics showing a relatively large difference between the first occasion respondents (Blue) and the second occasion respondents (Pink). For example, if H_3 was in fact the real situation, then assuming H_2, a bias of $B_3\ (H_2) = -.54$ would result in the estimation of the mean nights in Ontario; or on the other hand, if H_4 was assumed when H_3 is true, the resulting positive bias would then be equal to $B_3\ (H_4) = .73$.


Expenditures Per Party-Night (Y/X)
Blue
Pink
50
48
46.3
44
42
40
38
36
34
32
31
(42.01)
(33.38)
Expenditures Per Party (Y)
Pink 191.2
180
Blue 168.5
160
(172.8)
(186.0)
Mean Nights In Ontario (X)
Pink
Blue
6.15
6.0
5.5
5.0
4.5
4.11
4.0
3.64
(5.57)
.73 (+15%)
(-11%) -.54
(4.11)
0
0.27
1
1.5
2.0
a
H_1
H_2
H_3
H_4

5.2 Labour Force Survey (LFS)

A detailed description of the survey is given in a report by Platek and Singh (1976). Here, we shall briefly discuss the method of adjustment for non-response used in this survey and determine an idea about the magnitude of non-response bias. Also, since the survey consists of six rotation groups (panels) the households remaining in the sample for six consecutive months and then being replaced by another set of households, data on non-response rates by rotation groups may be utilized in determining the response probabilities.

Imputation for non-responding households are currently carried out as a combination of the procedures B and F. That is, firstly the labour force data for a non-responding household are imputed by copying data from the previous month's survey (with some update transformation) if the household responded for the previous survey. For the remaining non-responding households, imputation is carried out by using a balancing factor, which is the inverse of the 'response rate' at the balancing unit level. In the following table, the above procedure (B&F) is compared with procedure B alone. The proportions of the characteristics are given for the responding group and for the non-responding group for which data from the previous month were available.

The Table below gives Canada level data for certain labour force characteristics for two months, one of which when the non-response rates are usually the highest.

Table 1: Labour Force Data (Based on Unweighted Counts)

Char. Proportion \ Month	1		2	
	Response	Imputed	Response	Imputed
Employed	0.563	0.577	0.544	0.566
Unemployed	0.047	0.045	0.039	0.033
Not in LF	0.390	0.378	0.417	0.401
% of Imputed Households	-	6.91		1.74

For these months, approximately 4% of households classified as non-respondents were imputed by applying balancing factors. The remaining (6.91% and 1.74% in month 1 and 2 respectively) were imputed by copying the previous month's data. It should be noted that if the data imputed from the previous month is considered as the true data for the non-responding group in the present month, then because of the low non-response rates, the resulting biases would be negligible at the Canada level. For example, even in the case of "unemployed" in month 2 where differences between the two groups are noticeable, the relative overall bias resulting from the application of the balancing method on the 4% non-response would be of the order of .6%. Even if the balancing method was applied on the entire population of non-respondents (i.e. 5.74%), the relative overall bias would not exceed .9% .

However, at the provincial level, significant differences between the two sets of data were noted particularly for small provinces with higher non-response rates. Further work is being carried out by examining more characteristics at smaller levels over several months so as to account for seasonal variations and also for small area statistics.

Some other non-response data of interest from this survey are those of longitudinal data. With help of the Six Month Data File (SMDF) developed for the Labour Force Survey, it is possible to study the effect of successive interviews on the data as well as on the response rates by individual monthly surveys. We present below two tables, one showing the non-response patterns by number of interviews (or surveys) and the other showing the average household size by the frequency of responses over a six month period.

Table 2: Non-Response Patterns

Frequency of Response *	% of Total Number of Households
6	75.06
5	15.56
4	3.82
3	1.75
2	0.77
1	0.55
0	2.49
	100.00

* The frequency of response is the number of responses obtained over the six-month period.

Usually in panel surveys conducted on more than one occasion, there is a certain proportion of hard core non-respondents. If it is desirable to use the previous month's data for imputing the other non-respondents, then a combined procedure similar to that in the Labour Force Survey (LFS) may be applied.

For example, it would not be possible in the LFS to compensate for at least 2.49% of the non-respondents in any given survey month by using data from one of the previous survey months. Suitability of data imputed from any of the previous months will, however, depend upon the gross movements in the labour force characteristics, which can be studied using the same file.

Table 3: Average Household Size by Frequency of Response

Frequency of Response	1	2	3	4	5	6
Average Household Size	1.83	1.96	2.01	1.97	2.13	2.47

As expected, these figures show that smaller households are more difficult to contact. This suggests that for all characteristics having high correlation with the size of household, creation of post-strata, using household size as stratification variable, for compensation of non-response (procedure C) would be more efficient than uniform balancing factors (procedure B) used at design levels. Further, when such relationships are established with respect to various categories in the population, it would permit estimation of response probabilities for the corresponding categories. Knowledge of these response probabilities may be used in obtaining improved estimates. Also, such information could be useful in improving the procedure of data collection for the households which are difficult to contact, i.e. for those households whose response probability is low.

RESUME

Cet article a pour objectif d'étudier certaines des procédures utilisées pour atténuer les effets de la non-réponse. On élabore, à partir du concept des probabilités de réponse, un modèle simple qui analyse les erreurs de réponse et de non-réponse et on définit les divers éléments des deux types d'erreur dans le cadre de certaines procédures d'imputation. On présente également un graphique qui permet de visualiser l'importance du biais de non-réponse dans une situation donnée. On étudie enfin deux exemples pratiques d'imputation.

REFERENCES

[1] Chapman, David W. (1976), "A Survey of Non-Response Imputation Procedures", Report presented at the 1976 Annual Meeting of the American Statistical Association.

[2] Morgan, James M. and Sonquist, A. John (1963), "Problems in the Analysis of Survey Data, and a Proposal", Journal of the American Statistical Association, 58, 415-35.

[3] Platek, R., and Singh, M.P. (1976), "Methodology of the Labour Force Survey", Statistics Canada Publication.

[4] Politz, Alfred and Simmons, Willard (1949), "An Attempt to get the 'Not At Homes' into the Sample Without Callbacks", Journal of the American Statistical Association, 44, 9-31.

[5] Williams, W.H. and Mallows, C.L. (1970), "Systematic Biases in Panel Surveys Due to Differential Non-Response", Journal of the American Statistical Association, 65, 1339-49.

SOME ESTIMATORS OF POPULATION TOTALS FROM SIMPLE RANDOM SAMPLES CONTAINING LARGE UNITS

M.A.Hidiroglou and K.P. Srinath
Business Survey Methods Division

The problem considered is the estimation of population total of some characteristic from a simple random sample containing a few large or extreme observations. The effect of these large units in the sample is to distort the estimate of the population total. It is therefore important to correct the weights for such units or deflate their values at the estimation stage once they have been sampled and identified as unusually large units. In this paper, three estimators which alter the usual sampling weights have been considered. The efficiencies of these estimators have been worked out in terms of the ratio of the variance of the usual estimator of the population total to the mean square error of these estimators. An empirical study of these estimators is also discussed.

1. INTRODUCTION

The problem considered in this paper is the estimation of the population total of some characteristic from a simple random sample containing a few large or extreme observations. These observations are true observations belonging to the population that is being sampled. The presence of these observations in the sample will tend to make the usual estimate of the population total $\hat{Y}_o = N\bar{y}$ (where $\bar{y}$ is the sample mean and N the population size) exceed the population total Y by a considerable amount though the estimation procedure itself is unbiased. It is therefore important to deflate the weights for such units at the estimation stage once they have been sampled and identified.

Several techniques have been proposed to handle unusually large values. Tukey and McLaughlin (1963) considered trimmed and Winsorized sample means from symmetric distributions. Crow (1964) has studied weighting procedures for observations. Fuller (1960) studied one-sided Winsorized

means, Winsorization being applied to the largest observations only, assuming that the right tail of the distribution is well approximated by the tail of a Weibull distribution. Censored sample procedures have been considered by numerous authors (see for example Dixon (1960)). Searls (1966) proposed an estimator that used information external to the sample to predetermine a point, γ, which separates "large" sample observations from the rest.

Recently, in studying estimators for skewed populations, Jenkins, Ringer and Hartley (1973) have adopted biased estimators which were preferable to $N\bar{y}$. Their quadratic loss function incorporated both the squared bias and the variance of the estimators, i.e., the mean square error (MSE).

We confine our attention to estimators which involve only a change of the usual weights as this seems a realistic and practical approach in sample surveys. No knowledge of the number of large units (outliers) in the population is assumed. We propose three estimators which are designed to reduce the effect of these large observations. The efficiencies of these estimators are empirically investigated along with the efficiency of the post-stratified estimator which involves a knowledge of the number of outliers in the population. The criterion for comparison of the proposed estimators with the usual estimator $N\bar{y}$ is the ratio of the variance of the unbiased estimator to the mean square error of these estimators. It is shown that, in certain situations, these estimators will have a smaller mean square error than the usual estimator $N\bar{y}$.

2. THE ESTIMATORS

We assume that a population $\{Y_1, Y_2, \ldots Y_N\}$ of size N contains T large units. It is assumed that T is unknown. These outliers are elements of the population whose Y-value exceeds a prespecified value γ. A simple random sample of size n is drawn without replacement from the population and t outliers are identified. The estimators which we consider are:

$$\hat{Y}_1 = \sum_{i=1}^{t} y_i + \frac{N-t}{n-t} \sum_{i=t+1}^{n} y_i \quad . \tag{2.1}$$

$$\hat{Y}_2 = \frac{N}{n} \sum_{i=1}^{n} y_i - \frac{Nt}{2n} \left(\frac{n-t}{n}\right) \left(\sum_{i=1}^{t} \frac{y_i}{t} - \sum_{i=t+1}^{n} \frac{y_i}{n-t} \right) \tag{2.2}$$

and

$$\hat{Y}_3 = r \sum_{i=1}^{t} y_i + \frac{N-rt}{n-t} \sum_{i=t+1}^{n} y_i \quad . \tag{2.3}$$

Estimator (2.1) assigns weight one to the outlier units and adjusts the weights of the non-outliers so that the sum of the sample weights adds up to N. Estimator (2.2) assigns a weight to the outlier units which is dependent upon the number of outliers in the sample. Finally, estimator (2.3) generalizes estimator (2.1) in that it assigns an optimal weight r to the outlier and non-outlier units.

If T is known a priori, the post-stratified estimator is

$$\hat{Y}_4 = \frac{T}{t} \sum_{i=1}^{t} y_i + \frac{N-T}{n-t} \sum_{i=t+1}^{n} y_i \quad . \tag{2.4}$$

The bias and the mean square error (MSE) of these estimators are given in the following section.

3. THE MSE OF THE ESTIMATORS

We shall first consider the usual estimator of the population total $\hat{Y}_o$. $\hat{Y}_o$ may be expressed as the sum of outlier units and non-outlier units as:

$$\hat{Y}_o = \frac{N}{n} \{ \sum_{i=1}^{t} y_i + \sum_{i=t+1}^{n} y_i \} \quad . \tag{3.1}$$

The variance of $\hat{Y}_o$ in the form given in (3.1) is

$$V(\hat{Y}_o) = \{f^{-1}\ T(\frac{N-T}{N-1})\ (1-\delta)^2$$

$$+ N(f^{-1} - 1)\ \frac{T-1}{N-1}\ C_2^2\ \delta^2$$

$$+ N(f^{-1} - 1)\ \frac{N-T-1}{N-1}\ C_1^2\ \delta^2\}\ \bar{Y}_\nu^2 \qquad (3.2)$$

where f is the sampling fraction, δ is the ratio of the mean of the outlier units $\bar{Y}_\mu$ in the population to the mean of the non-outlier $\bar{Y}_\nu$ units in the population, C_1 and C_2 are the coefficients of variation for the non-outlier and outlier units in the population respectively.

It can easily be shown that the biases of $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$, for $T \geq 1$ are

$$B(\hat{Y}_1) = -T(1-f)(\delta-1)\ \bar{Y}_\nu \ , \qquad (3.3)$$

$$B(\hat{Y}_2) = \frac{-T(\delta-1)(N-T)\ \bar{Y}_\nu}{2N} \ , \qquad (3.4)$$

$$B(\hat{Y}_3) = -T(1-rf)(\delta-1)\ \bar{Y}_\nu \ . \qquad (3.5)$$

Note that estimators (2.1) and (2.3) are consistent whereas estimator (2.2) is not. The mean square error (MSE) of these estimators can be presented in two ways, depending on T. For T = 1, the mean square error can be derived exactly. For T > 1, the approximate MSE for $\hat{Y}_1$ and $\hat{Y}_3$ is obtained using $E(t) \doteq 1/E(t)$. For T > 1, the exact MSE for $\hat{Y}_2$ has been derived.

We first present the exact mean square errors associated with T = 1. Details of the derivations are not given here.

$$\text{MSE}(\hat{Y}_1) = \{(1-f)(1-\delta)^2$$

$$+ [\frac{f(N-1)}{n-1}(N-n) + N(1-f)(f^{-1} - \frac{N}{N-1})]\ C_1^2 \quad \bar{Y}_\nu^2\} \tag{3.6}$$

$$\text{MSE}(\hat{Y}_2) = \{\frac{N^2 n(1-f)}{f(N-1)(n-1)}[(1-f)(1 - \frac{n+1}{2n^2})^2 + 1]\ C_1^2$$

$$+ (1-\delta)^2\ [(1-f)[1 - \frac{f^{-1}}{2}(1 + \frac{1}{n})]^2 + f]\}\ \bar{Y}_\nu^2 \tag{3.7}$$

and

$$\text{MSE}(\hat{Y}_3) = \{[(1-f) + f(1-r)^2](1-\delta)^2$$

$$+ [\frac{f(N-r)^2}{n-1}\ \frac{N-n}{N-1} + N(1-f)(f^{-1} - \frac{N}{N-1})]\ C_1^2\}\ \bar{Y}_\nu^2. \tag{3.8}$$

The optimal value of r for (3.8) is given as

$$r_o = \frac{(1-f)\ C_1^2 + f(1-\delta)^2}{\frac{(1-f)}{N}\ C_1^2 + f(1-\delta)^2}\ .$$

Next, we provide expressions for MSE for $T > 1$.

$$\text{MSE}(\hat{Y}_1) \doteq \{(1-\delta)^2\ f(1-f)\ T\ (1 - \frac{T}{N}) + (T-1)\ f(1-f)\ C_2^2\ \delta^2$$

$$+ \frac{(1-f)}{f(N-T)}\ [(N-fT)^2 - f^2T]\ C_1^2 + T^2\ (1-f)^2\ (1-\delta)^2\}\ \bar{Y}_\nu^2 \quad . \tag{3.9}$$

$$\mathrm{MSE}(\hat{Y}_2) = \{[\frac{T(1-\delta)(N-T)(n-1)}{2n(N-1)}]^2$$

$$+ \left(\frac{C_2\delta}{2f}\right)^2 [Et + \frac{2Et^2}{n} + \frac{Et^3}{n^2}$$

$$- \frac{1}{T}(Et^2 + \frac{2Et^3}{n} + \frac{Et^4}{n^2})]$$

$$+ \left(\frac{C_1}{2f}\right)^2 [(4n - \frac{3\,Et^2}{n} - \frac{Et^3}{n^2})$$

$$- \frac{1}{N-T}(4n^2 - 4nEt - 3Et^2 + \frac{2Et^3}{n} + \frac{Et^4}{n^2})]$$

$$+ (\frac{1}{2nf})^2 (1-\delta)^2 V(nt + t^2)\} \quad \bar{Y}_\nu^2 \qquad (3.10)$$

where $V(nt + t^2) = n^2 V(t) + 2n \operatorname{Cov}(t,t^2) + V(t^2)$ and Et^k, $k=1,2,3,4$, are the moments obtained from the hypergeometric distribution given by

$$H(t|N,n,T) = \frac{\binom{N-T}{n-t}\binom{T}{t}}{\binom{N}{n}}, \quad 0 \le t \le T, \; N-T > n.$$

The mean square error of $\hat{Y}_3$ for $T > 1$ is

$$\mathrm{MSE}(\hat{Y}_3) \doteq \{r^2(1-\delta)^2 f(1-f)\, T(1 - \frac{T}{N}) + r^2(T - 1)\, f(1-f)\, C_2^2\, \delta^2$$

$$+ \frac{(1-f)}{f(N-T)} [(N-rfT)^2 - r^2f^2T]\, C_1^2 + T^2(1-rf)^2\, (1-\delta)^2\} \quad \bar{Y}_\nu^2 . \qquad (3.11)$$

The optimal value of r for $T > 1$ is obtained by minimizing (3.11). Differentiating (3.11) with respect to r and solving for r, we obtain

$$r_o = \frac{g_1(N,f,T,\delta,C_1)}{g_2(N,f,T,\delta,C_1,C_2)} \qquad (3.12)$$

where

$$g_1(N,f,T,\delta,C_1) = (1-\delta)^2 fT^2 + \frac{(1-f)TN}{N-T} C_1^2$$

and

$$g_2(N,f,T,\delta,C_1,C_2) = (1-\delta)^2 fT [(1-f)(1-\frac{T}{N}) + fT]$$

$$+ f(1-f)(T-1)[C_2^2 \delta^2 + \frac{T}{N-T} C_1^2].$$

The variance of the post-stratified estimator $\hat{Y}_4$ for $T > 1$ is given by

$$V(\hat{Y}_4) \doteq \{C_1^2 [f^{-1}(1-f)(N-T) + \frac{T}{nf}]$$

$$+ C_2^2 \delta^2 [f^{-1}(1-f)T + \frac{N-T}{nf}]\} \bar{Y}_\nu^2 . \qquad (3.13)$$

4. AN EMPIRICAL INVESTIGATION OF THE ESTIMATORS

To investigate the efficiency and utility of the proposed estimators, we have used a variety of artificial populations. We have studied the relative efficiency of these estimators for various values of C_1, C_2, δ, f, N and T. The relative efficiency is defined as the ratio of the variance of the usual estimator of the total $\hat{Y}_o$ to the mean square error of $\hat{Y}_i$, i=1,2,3,4. The empirical investigation has been extensive and in view of the difficulty of presenting a great number of tables, only six tables are presented. Tables 1 through 5 are constructed to reveal a difference in the behaviour of the estimators $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$ for various values of C_1, C_2, δ, f and T for a given value of N. Within

each of these tables C_2 and T vary while δ, f and C_1 are fixed. The tables differ from each other by having one of the variables δ, f or C_1 vary while the other two variables are fixed. Table 6 differs from the others in that a large value of N and a small sampling fraction f have been used. The conclusions drawn from these tables, in general, should apply to other populations.

Tables of Relative Efficiencies

Estimators			$\hat{Y}_1$		$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$	
1.			δ=5	f=0.3		C_1= 0.5	N=500	
T \ C_2	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0
2	1.26	-	1.16	-	1.26 (1.10)	-	0.62	-
4	1.37	2.32	1.41	2.00	1.41 (1.49)	2.32 (0.78)	0.77	0.49
10	1.02	2.13	1.30	2.08	1.37 (2.15)	2.17 (1.32)	1.03	0.74
15	0.75	1.69	1.10	1.89	1.30 (2.43)	1.92 (1.62)	1.15	0.84
25	0.48	1.18	0.81	1.56	1.20 (2.72)	1.61 (2.02)	1.28	0.94
80	0.14	0.40	0.35	0.83	1.06 (3.12)	1.20 (2.77)	1.43	1.07
2.	δ=5	f=0.1		C_1=0.5			N=500	
2	1.37	-	1.28	-	1.37 (1.19)	-	0.40	-
4	1.75	3.22	1.56	2.13	1.75 (1.76)	3.22 (0.75)	0.53	0.30
10	1.85	4.17	1.85	2.63	2.04 (3.22)	4.17 (1.46)	0.78	0.53
15	1.53	3.70	1.85	2.63	1.96 (4.11)	3.84 (2.00)	0.92	0.65
25	1.06	2.78	1.67	2.00	1.69 (5.34)	3.12 (2.87)	1.09	0.79
3.	δ=10	f=0.3		C_1=0.5			N=500	
2	1.78	-	1.72	-	1.78 (1.16)	-	0.48	-
4	1.64	3.12	1.78	2.56	1.78 (1.58)	3.12 (0.89)	0.75	0.46
10	0.92	2.04	1.30	2.13	1.45 (2.24)	2.17 (1.47)	1.15	0.76
15	0.64	1.51	1.02	1.82	1.31 (2.51)	1.85 (1.79)	1.30	0.87
25	0.40	0.99	0.71	1.41	1.19 (2.78)	1.51 (2.18)	1.45	0.98
80	0.12	0.33	0.31	0.71	1.06 (3.14)	1.16 (2.86)	1.59	1.10

4.	δ=10		f=0.1		C_1=0.5		N=500	
2	2.43	-	1.92	-	2.43 (1.23)	-	0.27	-
4	3.03	6.25	2.32	2.94	3.12 (1.89)	6.25 (0.87)	0.46	0.27
10	2.08	5.00	2.27	2.94	2.56 (3.48)	5.26 (1.70)	0.82	0.53
15	1.51	3.70	2.08	2.78	2.13 (4.41)	4.17 (2.30)	1.00	0.66
25	0.94	2.44	1.69	2.50	1.72 (5.66)	3.03 (3.28)	1.20	0.81
5.	δ=5		f=0.3		C_1=1.0		N=500	
2	1.06	-	1.06	-	1.06 (1.16)	-	0.83	-
4	1.12	1.47	1.12	1.39	1.14 (1.53)	1.47 (0.80)	0.88	0.63
10	1.00	1.67	1.14	1.64	1.18 (2.17)	1.69 (1.33)	1.02	0.79
15	0.81	1.52	1.05	1.64	1.16 (2.44)	1.64 (1.64)	1.10	0.87
25	0.54	1.14	0.84	1.45	1.14 (2.73)	1.49 (2.03)	1.20	0.95
80	0.16	0.41	0.37	0.83	1.06 (3.12)	1.19 (2.77)	1.37	1.06
6.	δ=5		f=.01		C_1=0.5		N=10,000	
5	1.06	1.19	1.05	1.14	1.07 (2.88)	1.19 (1.09)	0.52	0.23
15	1.22	1.64	1.16	1.41	1.22 (6.41)	1.64 (2.39)	0.57	0.29
25	1.33	2.04	1.25	1.64	1.35 (9.75)	2.04 (3.70)	0.62	0.35
25	1.51	2.70	1.41	1.96	1.54(15.79)	2.70 (6.22)	0.70	0.45
65	1.61	3.12	1.54	2.17	1.67(21.08)	3.12 (8.61)	0.77	0.52
85	1.61	3.33	1.61	2.32	1.75(25.76)	3.45(10.89)	0.83	0.58

Note: Dashes indicate that C_2 is non-existent for these cases. The numbers in brackets are the optimal r_o values given by (3.12).

It is seen from the above tables that, for fixed δ, f, C_1, C_2, and N, the efficiencies of the estimators decrease after an initial improvement as T increases. The efficiency gain in using these estimators increases as the coefficient of variation C_2 of the outlier units increases. Comparing the values in Table 1 with those in Table 5, we see that as C_1 increases, the efficiencies of the estimators decrease for small values of T and increase after a

certain number of outliers has been reached. Comparing values in Tables 1 and 3, we see that as δ increases from 5 to 10, gains in efficiency are not uniform. In fact, for large T, there is a greater loss in efficiency. This is due to the fact that the bias term of the estimators dominates the mean square error as δ increases. Referring to Tables 1 and 2, 3 and 4, it is seen that as f decreases, gains in efficiencies of the estimators increase.

To stress the effectiveness of these estimators, a fairly large population of N=10,000 and a small sampling fraction of f=0.01 have been used. The results are given in Table 6. Note that for a few number of outliers in the population, the gain in using these estimators is not very considerable. However, as the number of outliers in the population increases, the effectiveness of these estimators improves quite significantly.

It is possible to make the following general observations. The best estimator to use with respect to efficiency is $\hat{Y}_3$. $\hat{Y}_2$ has lower efficiency than $\hat{Y}_1$ for a small number of outliers, however, after a certain number of outliers has been reached, $\hat{Y}_2$ is superior to $\hat{Y}_1$. Hence, $\hat{Y}_2$ is to be preferred to $\hat{Y}_1$ in the presence of a moderate number of outliers. For a small number of outliers, the post-stratified estimator $\hat{Y}_4$ is not as good as the other estimators because the allocation between the post-strata is likely to be poor, being very different from the optimum allocation in such cases. But, as expected, once a certain number of outliers is reached, it is superior to all estimators including $\hat{Y}_o$.

$\hat{Y}_3$, the optimal estimator, requires a knowledge of T, C_1, C_2, and δ. One way of determining r_o is to estimate T, C_1, C_2 and δ from the sample and use these in the expression (3.12). Estimating r_o using sample values could imply a departure from optimal efficiency of $\hat{Y}_4$. To study this possible departure, the efficiency of $\hat{Y}_3$ has been investigated for different values of r_o $(1+\Delta)$, where $0.0 \leq \Delta < 1.0$. Two situations have

been investigated. The first one being a large population size of 10,000 with an associated small sampling fraction of 0.01 and the second being a small population size of 500 with a fairly high sampling fraction of 0.3. The results are given in Tables 7 and 8.

Tables of Relative Efficiencies of $\hat{Y}_3$ for $r_o(1+\Delta)$

7.	$\delta=5$	$f=0.3$	$C_1=0.5$		$C_2=1.0$	$N=500$
T	2	4	10	15	25	80
r_o / $\|\Delta\|$	1.10	1.49	2.5	2.43	2.72	3.12
0.0	1.26	1.41	1.37	1.30	1.20	1.06
0.1	1.26	1.41	1.35	1.26	1.15	0.94
0.2	1.26	1.40	1.31	1.19	1.04	0.69
0.3	1.26	1.38	1.23	1.09	0.89	0.47
0.4	1.25	1.35	1.15	0.97	0.74	0.33
0.5	1.25	1.32	1.05	0.85	0.61	0.24
0.6	1.23	1.29	0.95	0.74	0.51	0.18
0.7	1.23	1.25	0.86	0.64	0.42	0.14
0.8	1.21	1.21	0.77	0.56	0.35	0.11
0.9	1.20	1.16	0.69	0.48	0.29	0.09

8.	$\delta=5$	$f=0.01$	$C_1=0.5$		$C_2=1.0$	$N=10,000$
T	5	15	25	45	65	85
r_o / $\|\Delta\|$	2.88	6.41	9.75	15.79	21.08	25.76
0.0	1.069	1.216	1.346	1.546	1.678	1.755
0.2	1.069	1.216	1.345	1.545	1.673	1.746
0.4	1.069	1.216	1.344	1.541	1.541	1.664
0.6	1.069	1.216	1.343	1.535	1.648	1.697
0.8	1.069	1.215	1.341	1.527	1.626	1.656

From the preceding tables, it is seen that when there is a low number of outliers, the efficiency of $\hat{Y}_3$ is not significantly affected by departures from optimal r_o. As the number of outliers increases in the first (N=500) population, even small departures from optimal r_o result in low efficiency. Note that in the case of the second population (N=10,000), departures from optimal r_o are not significant even for large number of outliers in the population.

5. CONCLUSIONS

When the sampling fraction f and the number of outliers T are small, use of the estimator $\hat{Y}_1$ would result in substantial gains in efficiency. If f and T are moderately large, use of $\hat{Y}_2$ is recommended. $\hat{Y}_3$ can be used to advantage if values of C_1, C_2, δ and T are approximately known from previous surveys. Deviations from the optimal r_o associated with $\hat{Y}_3$ will not affect the efficiency if T is small. If T is large and known, it is obvious that the post-stratified estimator $\hat{Y}_4$ should be used.

RESUME

On veut obtenir une estimation de la valeur totale d'une caractéristique donnée dans une population à partir d'un échantillon aléatoire simple comportant quelques observations extrêmes ou très grandes. La présence de grandes unités de ce genre dans l'échantillon fausse l'estimation de la valeur totale. Si donc après échantillonnage, on repère des unités extrêmes, il importe d'en adjuster la pondération ou d'en corriger la valeur au stade de l'estimation. Dans cet article, on examine trois estimateurs qui modifient la pondération d'échantillonnage habituelle. L'efficacité de ces estimateurs a été établie en fonction du rapport de la variance de l'estimateur habituel du chiffre total à l'erreur quadratique moyenne de ces estimateurs. Une étude empirique a été effectuée, basée sur cette theorie.

ACKNOWLEDGMENT

The authors would like to thank Dr. M.P. Singh and the referee for useful comments.

REFERENCES

[1] Bershad, M., "Some Observations On Outliers", Unpublished dittoed memorandum, 1960, Statistical Research Division, U.S. Bureau of Census.

[2] Chinnappa, N., "A Preliminary Note On Methods of Dealing With Unusually Large Units In Sampling from Skew Populations", Unpublished, IASMD technical memorandum, February 1976.

[3] Crow, Edwin L., "The Statistical Construction of a Single Standard from Several Available Standards", IEEEE Transactions On Instrumentation and Measurement, 13 (1964), 180-5.

[4] Dixon, W.J., "Simplified Estimation From Censored Normal Samples", Annuals of Mathematical Statistics, 31 (1960), 385-91.

[5] Fuller, W.A., "Simple Estimation for the Mean of Skewed Populations", 1960, U.S. Bureau of Census.

[6] Hartley, Herman O., and Rao, J.N.K., "A New Estimation Theory for Sample Surveys", Biometrika, 55 (March 1968), 547-57.

[7] Jenkins, O.C., Ringer, L.G., and Hartley, H.O., "Root Estimators", Journal of the American Statistical Association, 68 (1973) 414-19.

[8] Rao, C.R., "Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations", in Foundation of Statistical Inference. Holt, Rinehart and Winston of Canada Ltd., (1971), 177-202.

[9] Searls, D.T., "An Estimator Which Reduces Large True Observations", Journal of the American Statistical Association, (1966), 1200-4.

[10] Tukey, J.W., and McLaughlin, D.H., "Less Vulnerable Confidence and Significance Procedures for Location Based On A Single Sample: triming/Winsorization 1", Sankhya, Series A, 25 (1963) 331-52.

REDESIGN OF THE SOUR CHERRY, PEACH AND GRAPE OBJECTIVE YIELD SURVEYS IN THE NIAGARA PENINSULA

G. Davidson
Institutional and Agriculture Survey Methods Division

Objective yield surveys have been conducted annually in the Niagara Peninsula since 1964. The aim of each of these annual surveys is to provide a forecast of the marketable production change in the region from the previous year. These estimates are determined far enough in advance of the harvest to enable them to serve as important factors in price negotiations between growers and processors, as well as indicators of particular crop situations which could necessitate immediate changes in strategy by the marketing agencies. In 1973 an extensive redesign project was initiated. This report provides a summary of the sample design, data collection procedures and estimation procedures which were incorporated in the redesign of the sour cherry, peach and grape objective yield surveys.

1. INTRODUCTION

The objective yield surveys which were conducted prior to 1974 employed list frames which, unfortunately, could not be adequately updated from year to year. Thus, changes in land use in the Niagara Peninsula (see Appendix A for a table which illustrates the land use changes), which resulted in diminished sample sizes for each survey, could not be accounted for with the existing out-of-date lists. This shortcoming formed the basis for the decision in March, 1973 to switch from a list frame based sample design for each survey to a common area frame based sample design for the three surveys.

The development of a common area frame for the three surveys would provide a reliable multipurpose frame from which a sample of blocks (i.e. contiguous areas of trees and/or vines), for each of the three

surveys could be selected. Also, the common area frame based sample design would easily allow for the introduction of sample rotation or a sample increase in subsequent years. In addition, with this sample design the sample of blocks should be more clustered than with the list frame based design, which would mean a saving in terms of travelling time and costs.

The total marketed production (i.e. the total amount of fruit which actually was sold either as fresh fruit or to processors) figures released each year by the Ontario Fruit and Vegetable Statistics Committee are reported in weight units. However, each of the three objective yield surveys employs a ratio of change estimator which is the ratio of the estimated total number of pieces of marketable fruit in the region in the current year to the corresponding estimated total for the previous year. Also, as the estimates are produced prior to the fruit being harvested, they do not account for the change in the number of pieces of marketable[1] fruit between the time of the survey and harvest time. Therefore, unless there is a consistent loss of fruit each year between the time of the survey and harvest time, then the accuracy of each ratio of change estimator would be affected. Thus, each objective yield survey estimator supposes the following two major assumptions:

(1) There is a very high positive correlation between fruit counts and fruit weights.

(2) Each year there is a consistent loss of fruit between the time of the survey and harvest time (including fruit loss at harvest time due to mechanical harvesting).

Hence, it was also decided in March of 1973 that research work should be initiated in the area of fruit estimation procedures at the tree, vine and sample block levels. This research work would examine the validity of the two major assumptions supposed by each survey estimator, as well as investigate such problems as forecasting the weight gain of each fruit between survey time and harvest time, and determining the effects of fruit variety on fruit yield.

---

[1] The term 'marketable' refers to the total amount of fruit grown which is capable of being sold either as fresh fruit or to processors.

2. SAMPLE DESIGN

2.1 Outline of Design

The sample design which was developed in 1973-74 may be divided into an agricultural EA[2] component and a non-agricultural EA component. The agricultural EA component employs a four stage sample design to obtain a sample of blocks for each survey. The non-agricultural EA component employs a single stage sample design to obtain a sample of non-agricultural EA's. The design stages for each component of the sample design are illustrated by the flow diagram in Figure 1.

2.2 The Area Frame

The universe for the area frame comprises the following municipalities as defined by the 1971 Census of Agriculture:

(1) Within Niagara Regional Municipality--the municipalities of Lincoln West, Niagara Falls, St. Catharines, Welland, Grimsby, Lincoln, Niagara-on-the-Lake, Pelham and Thorold.

(2) Within Wentworth County--the municipality of Saltfleet.

The area defined above contains 210 agricultural EA's and 179 non-agricultural EA's (excluding urbanized core EA's).

2.3 Stratification of EA's

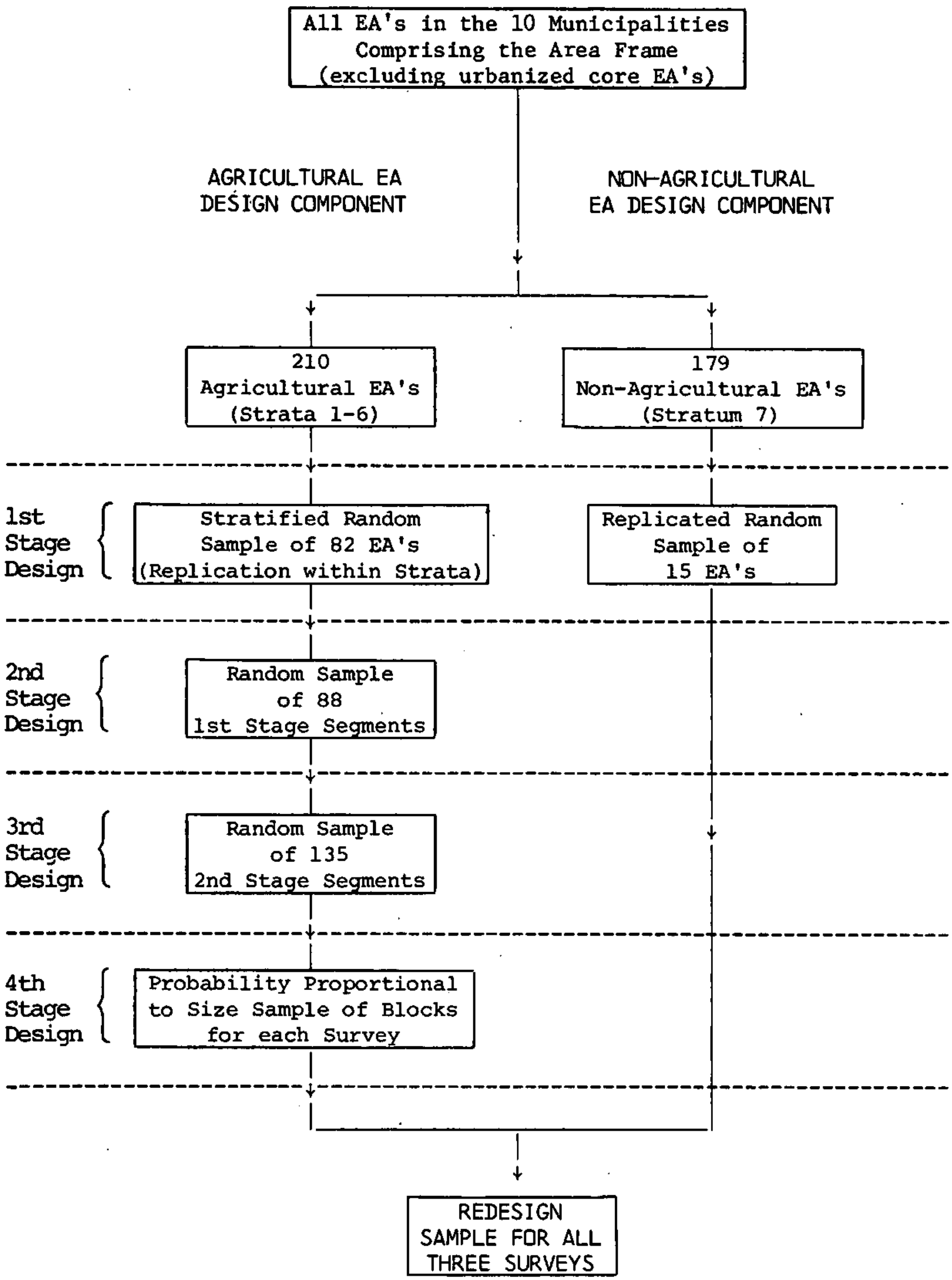
The 210 agricultural EA's were stratified into 6 strata on the basis of 1971 Census of Agriculture data pertaining to the total acreage in the EA which was devoted to growing each of the four fruits (i.e. sour

---

[2]An EA, or enumeration area, is the area canvassed by one enumerator in collecting the 1971 Census of Population data. An agricultural EA is an EA which contains the headquarters of at least one census-farm (i.e. a farm, ranch or other agricultural holding of one acre or more with sales of agricultural products, during the 12-month period prior to the census, of $50 or more).

SAMPLE DESIGN FLOW DIAGRAM

Figure 1



cherries, peaches, grapes and pears[3]). The 179 non-agricultural EA's were assigned to stratum 7.

2.4 Agricultural EA Design Component

(a) First Stage Design - Selection of A Replicated Random Sample of EA's Within Strata

From the EA's assigned to each of the 6 agricultural EA strata, a replicated random sample of EA's was selected. The number of agricultural EA's to be sampled from each stratum was dependent upon the time, manpower and monetary resources available for the ground mapping phase of the redesign project which was conducted in October, 1973. After consideration of these factors it was decided to select a sample of 82 agricultural EA's. The number of strata, the stratum boundaries, the sequential order of allocating the agricultural EA's to strata, and the number of agricultural EA's to be sampled from each stratum were determined by testing various combinations of these 4 factors and evaluating the results of each design. A total of 50 different designs were tested with the design chosen emerging as the most efficient[4] of the designs tested. As independent replicates were selected within a particular stratum, EA's could be selected in more than one replicate from that stratum. This resulted in 65 distinct EA's being selected in the total sample of 82 agricultural EA's. Table 1 indicates the stratum boundaries, the number of agricultural EA's in the population in each stratum, the agricultural EA sample size in each stratum, the number and size of replicates selected in each stratum and the sampling fraction for each agricultural EA stratum.

---

[3] The area frame based sample design was developed to allow for the introduction of a pear objective yield survey at a later date. Funds for this survey have not yet been allocated, although sample pear tree measurements involving 20 sample trees have been made each year since 1974.

[4] The efficiency of the different designs was measured by calculating the relative efficiency for each of the four fruits of interest. The relative efficiency for an item (in %) is defined as

$$\left[\frac{\text{Variance for item using simple random sample}}{\text{Variance for item using stratified design}}\right] \times 100$$

with the same overall sample size being used to calculate both variances.

TABLE 1

Stratum Number	Stratum Boundaries	Number of Agricultural EA's in the Population in the Stratum	Agricultural EA Sample Size in the Stratum	Number of Replicates in Sample from the Stratum	Number of Agricultural EA's per Replicate in Sample from the Stratum	Sampling Fraction for the Stratum
1	Agricultural EA's containing no sour cherries, peaches, grapes or pears (i.e. the 'zero' stratum).	40	8	4	2	1:5.0
2	Agricultural EA's containing 400 or more acres of grapes.	15	12	4	3	1:1.25
3	Agricultural EA's containing 100 or more acres of peaches <u>and</u> which were not previously allocated to stratum 2.	21	20	5	4	1:1.05
4	Agricultural EA's containing 15 or more acres of sour cherries <u>and</u> which were not previously allocated to strata 2 or 3.	22	15	5	3	1:1.47
5	Agricultural EA's containing 100 or more acres of grapes <u>and</u> which were not previously allocated to strata 2, 3 or 4.	19	6	3	2	1:3.17
6	All remaining agricultural EA's.	93	21	7	3	1:4.43
	TOTALS	210	82	-	-	1:2.56

One reason for the very large sampling fractions for strata 2, 3 and 4 is that the acreages devoted to growing each of the four fruits of interest in the EA's allocated to these three strata amount to a very high percentage of the total acreage in the universe devoted to growing these four fruits. This fact is evident from Table 2, which utilizes 1971 Census of Agriculture data.

TABLE 2

Stratum Number	Number of Agricultural EA's in the Population Allocated to the Stratum	Percentage of the Total Acreage in the Universe Devoted to Growing Each Fruit Accounted for by the Agricultural EA's Allocated to the Stratum			
		Peaches	Sour Cherries	Grapes	Pears
2	15	25.6	24.1	50.6	32.8
3	21	61.8	40.1	15.4	31.2
4	22	5.7	25.5	9.7	17.8
TOTALS	58	93.1	89.7	75.7	81.8

(b) Second Stage Design - Selection of First Stage Segments Within EA's

As agricultural EA's are relatively large in size, it was necessary to divide up each of the selected agricultural EA's into a number of primary, or first stage, segments. This segmenting of EA's was accomplished by utilizing topographical maps, 1971 Census maps containing the EA boundaries, 1972 infra-red aerial photographs of the region and by adhering to the following two criteria whenever possible:

(i) The agricultural EA should be divided up into a number of first stage segments which possess approximately equal amounts of agricultural activity (i.e. the first stage segments should all have approximately the same land area devoted to the growing of crops, and in particular the growing of fruit crops, as determined from the infra-red aerial photographs).

(ii) The boundaries of each first stage segment should be natural or evident man-made boundaries such as streams, rivers, roads, ditches, etc.

Whenever a conflict arose between these two criteria it was usually decided that criterion (ii) should take priority over criterion (i) in view of the difficulties that had been experienced by enumerators in accurately locating artificial or imaginary boundaries such as property lines, township lines, etc. Then, for 76 of the total sample of 82 agricultural EA's, one first stage segment was randomly selected from each EA. For each of the remaining 6 agricultural EA's, two first stage segments were randomly selected.

(c) Third Stage Design -
Selection and Ground Mapping of Second Stage Segments Within First Stage Segments

Since the first stage segments selected were much too large in area for an enumeration group to cover, each selected first stage segment was divided up into a number of second stage segments. To enable segmentation of the selected first stage segments to be carried out,double size enlargements (scale equals 660 feet to the inch) of 1972 aerial photographs of the selected first stage segments were obtained. These enlargements were necessary as the infra-red aerial photographs lacked sufficient detail to be of any help in this segmentation. Utilizing these enlargements and the criteria previously described for segmenting the selected agricultural EA's, the second stage segmentation was completed. Then, from each of the 88 selected first stage segments, one second stage segment was randomly selected.

With the selection of the second stage segments completed, the ground mapping phase of the redesign project was carried out in October of 1973. The main objectives of this phase were:

(1) To make initial contact with the owner(s) and/or operator(s) of all land contained within each selected second stage segment and solicit their co-operation in the surveys.

(2) To find out which type(s) of fruit were contained in each selected second stage segment, and to draw a map of each segment indicating where the trees or vines were located.

(3) To determine how many vines or trees of interest were contained within each selected second stage segment.

In order to simplify the task of locating the second stage segments on the ground, each enumeration group was provided with topographical maps which indicated the location of each of their assigned second stage segments in relation to nearby roads, towns, etc. Also, for each assigned second stage segment, an enumeration group received a folder containing the necessary forms and an aerial photograph enlargement of the segment and the surrounding land.

The ground mapping of the 88 selected second stage segments was completed in less time than anticipated, in part due to the abnormally good October weather that was experienced. As a result, since all of the necessary aerial photographs were in St. Catharines, it was decided that from each selected first stage segment belonging to an agricultural EA sampled from strata 2, 3 or 4 (i.e. the three strata which together contain most of the total acreage planted in the fruits of interest), an additional second stage segment would be randomly selected without replacement and ground mapped. The number of additional second stage segments selected in St. Catharines was 47, thus bringing the total number of second stage segments which were ground mapped to 135.

In order to provide some indication of the actual percentage of the total number of trees and vines in the peninsula which were contained in the sample of second stage segments, the tree and vine space counts obtained during ground mapping were totaled for all selected second stage segments. These totals are compared in Table 3 with the 1971 Fruit Tree Census figures for the Niagara district compiled by the Ontario Ministry of Agriculture and Food, see [5] and [6].

TABLE 3

Fruit Type	Total Number of Trees or Vine Spaces in all Selected Second Stage Segments	1971 Fruit Tree Census Total*	Percentage (1) is of (2)
	(1)	(2)	
Grapes	469,108	8,221,388	5.7
Peaches	44,176	738,713	6.0
Sour Cherries	12,953	197,551	6.6
Pears	29,106	417,235	7.0

*These figures do not represent 100% of the trees and vines in the Niagara district. The amount of grower non-response is given in the report at the provincial level only so that determining the amount of non-response for the Niagara district alone is not possible.

(d) Fourth Stage Design - Selection of Blocks Within Second Stage Segments

During the ground mapping phase each selected second stage segment was divided up into a number of blocks with block boundaries being natural or evident man-made boundaries whenever possible. The block was felt to be the largest part of a selected second stage segment which an enumeration group could cover each year, and from which sufficiently accurate producing tree or vine space (i.e. a space in a vineyard which does contain or could contain a vine) counts could be obtained. The size of the blocks varied within each selected second stage segment with some blocks containing only one orchard or vineyard, while others contained 3 or 4 small orchards or vineyards. The block served as the basic reporting unit for the tree and vine space counts obtained during the ground mapping phase, although within block counts for orchards and vineyards were also reported for almost all segments.

The selection of a block (or blocks) from each sample second stage segment was carried out back in Ottawa following the ground mapping work. For each of the three fruits (i.e. sour cherries, peaches and grapes)

present in a second stage segment, the selection method consisted of selecting a sample block (selecting with replacement if two or three of the fruits were present in at least one of the blocks in the segment) with probability proportional to the number of vine spaces or trees of that particular fruit reported in a block in the segment.

Due to the absence of fruit, and in many cases leaves, on the trees in October during the ground mapping phase, the ability to accurately determine the number of producing trees in an orchard was greatly reduced. Therefore, total tree counts were obtained in each orchard and no attempt was made to try and obtain separate counts of producing and non-producing trees. Also, the enumerators estimated the number of vine spaces in each vineyard instead of estimating the total number of vines, due to the difficulty in obtaining an accurate estimate of the number of vacant vine spaces without spending too much time walking up each pair of rows in the vineyard.

2.5 Non-Agricultural EA Design Component

(a) First Stage Design - Selection of a Replicated Random Sample of EA's from Stratum 7

From the 179 non-agricultural EA's (excluding urbanized core EA's) assigned to stratum 7, five independent replicates of EA's, each consisting of three EA's, were randomly selected. These 15 sample EA's were located in or near urban areas and, thus, the relatively small size of each EA meant that checking the entire EA for signs of bearing vines or trees of interest could easily be completed.

3. DATA COLLECTION

3.1 General

The sour cherry survey is conducted each year during the latter part of June while the peach and grape surveys are conducted concurrently each year during the latter part of July. The data collection phase of the sour cherry survey is usually completed in 5 days while for the peach and grape surveys 8 or 9 days are usually required. Following each day's work the field supervisor manually edits the forms completed that day. Then the edited data for that day is entered, via a remote computer terminal located in the field supervisor's hotel room, into a computer located in Toronto. Preliminary tabulation runs can then be performed and evaluated each night and the data files created can also be accessed by personnel back in Ottawa. The use of this computer terminal in the field has meant substantial data processing time savings, and has enabled final ratio of change estimates to be released to users within 48 hours of the completion of field work.

The remainder of this section highlights some of the field procedures which constitute the data collection phase.

3.2 Sour Cherry and Peach Surveys

From each block selected in the fourth stage of the agricultural EA design component which contains one or more sour cherry or peach orchards, a sample of 4 producing sour cherry or peach trees was randomly selected without replacement in the summer of 1974. In order to take advantage of the expected high positive correlations between measurements from the same tree on successive occasions, the enumerators return to the same sample trees each summer unless a sample tree has died or has been pulled up.

In 1974, for the sour cherry survey a sample limb (or limbs) was selected from each sample tree employing a branch sampling method described by Jessen [3]. This method consists of: (i) selecting a primary limb (i.e.

a main limb originating at the first branching of the trunk of a tree) with probability proportional to the cross-sectional area (i.e. CSA) of a primary limb, and (ii) then proceeding out along the selected primary limb, utilizing this probability proportional to size limb selection procedure to select a smaller limb at each branching point of this primary limb, until a sample limb, or count limb, of the desired size has been selected. For this survey the desired size for a count limb is a limb whose CSA is greater than or equal to 5% but less than or equal to 15% of the primary limbs cumulative CSA total. However, it was not always possible to select a count limb on a sample sour cherry tree which satisfied this size criterion. In such instances two count limbs were selected following specified selection procedures. After selecting the count limb(s) on a sample tree, the enumerator painted a yellow strip around each limb near its base for future identification purposes. Each summer the enumerators locate the count limb(s) on each sample tree, unless the limb(s) has been sawn off, and count all of the sour cherries, excluding culls (i.e. immature or damaged fruit which will not be harvested), on each count limb.

Each year in the peach survey the enumerators count all of the peaches, excluding culls, on each sample tree. It would not have been feasible to select only a sample limb (or limbs) on each selected peach tree for fruit counting purposes due to two main reasons:

(i) the presence in the region of the disease peach canker, which spreads rapidly unless the infected limbs are sawn off, and

(ii) generally, the fruit is distributed in a much more haphazard manner on a peach tree than on a sour cherry tree (e.g. an entire side of a producing peach tree may contain no marketable peaches, or a large limb may contain only a few peaches while a much smaller limb may be heavily laden with fruit). As a result of this nonuniform growth characteristic exhibited by peach trees, the correlation between limb cross-sectional areas and fruit counts would be much lower for peach trees than for sour cherry trees.

Also, each year it is necessary for the enumerators to determine the number of producing sour cherry or peach trees in each selected block.

3.3 Grape Survey

From each selected block which contains one or more vineyards a sample of 5 producing vines was randomly selected without replacement in the summer of 1974. As in the sour cherry and peach surveys, the enumerators return to the same sample vines each summer unless a sample vine has died or has been pulled up.

Each year the enumerators count the bunches on each sample vine, excluding bunches consisting of 5 or fewer berries, and then randomly select 5 of these bunches without replacement on each vine. Next the number of berries, excluding culls, is determined for the 5 bunches selected on each sample vine. Finally, the enumerators estimate the number of vine spaces in each selected block.

3.4 Second Stage Segments Not Containing Any Marketable Fruit of Interest

During the ground mapping phase in October of 1973 a number of selected second stage segments (in addition to selected second stage segments belonging to agricultural EA's in the 'zero' stratum) were found not to contain any bearing sour cherry or peach trees nor bearing vines. This situation had been expected to arise due to: (i) the three fruits surveyed were not distributed uniformly throughout the selected agricultural EA's and, therefore, it was impossible to ensure that every first and second stage segment contained bearing vines or bearing trees of interest, and (ii) the random selection procedures which were employed to obtain a sample second stage segment (or segments) within each selected agricultural EA. As a result, each summer a visit is made to approximately one-third of these second stage segments. Each of the segments visited each year is checked thoroughly by the enumerators for vines or trees of interest which have been planted since the last visit to that segment. If it is found that a segment now contains enough

producing vines or trees of interest to warrant the selection of a sample, then the field procedures for that survey are followed and the necessary measurements are taken.

3.5 Selected Non-Agricultural EA's

The initial visit to each of the 15 non-agricultural EA's selected was made in the summer of 1974. At this time the enumerators covered these EA's thoroughly and found that each selected EA did not contain any bearing vines or trees of interest. Furthermore, the enumerators indicated that the prospect of future plantings of trees or vines in any of these EA's appeared to be quite remote. As a result, these selected non-agricultural EA's will not be revisited until 1978 or 1979,as vines or trees of interest subsequently planted in any of these EA's would not be of bearing age until 1978 at the earliest.

4. UPDATING THE REDESIGN SAMPLE IN 1976

As a means of allowing for the introduction of newly bearing trees and vines into the sample, it was originally decided that, beginning in 1976, the sample of EA's in one or two replicates from each stratum would be rotated out of the sample of EA's and new replicates rotated into the sample. This planned sample EA rotation scheme would have meant that by 1980 all of the originally selected replicates in each stratum would be rotated out of the sample of EA's. However, due to budgetary limitations this planned 20% rotation of sample EA's was not carried out in 1976, nor will it be introduced into the 1977 surveys. Instead, in 1976 sufficient funds were made available to permit only the following additions to the sample:

(i) A sixth replicate of agricultural EA's (i.e. 4 agricultural EA's) was randomly selected from all agricultural EA's belonging to stratum 3. Then, following the sample design procedures previously outlined, a total of 4 first stage segments and 8 second stage segments were randomly selected.

(ii) Two second stage segments were added in order to replace two second stage segments which had to be dropped from the surveys because the owners of the land in these 2 segments refused to participate in the surveys. Each second stage segment added was randomly selected without replacement from the same first stage segment which contained the refusal segment.

In June, 1976 each of these 10 additional second stage segments was ground mapped and divided up into a number of blocks. Then, from each second stage segment a sample block was selected with probability proportional to size for each of the three fruits of interest found in the segment. From these 10 second stage segments a total of 7 sample blocks were selected for the sour cherry survey, 6 for the peach survey and 4 for the grape survey. The selection of these sample blocks for each survey helped to offset a reduction between 1974 and 1976 in the number of sample blocks included in each survey. This decrease in the number of sample blocks amounted to 6 sample blocks in the grape survey, 7 in the sour cherry survey and 8 in the peach survey. These decreases (excluding those resulting from the two second stage segment refusals) were the result of agricultural land going out of production, or a change occurring in the crop being grown on the land. Also, these decreases constituted the reason for adding a sixth replicate of agricultural EA's to the sample of EA's from stratum 3 instead of rotating out one of the existing replicates and rotating in a new replicate.

5. SUMMARY

The three redesigned objective yield surveys became operational in the summer of 1974, however, as the ratio of change estimator employed requires two consecutive years of data, estimates utilizing the redesign sample could not be obtained until the summer of 1975. As a result, it was necessary to conduct the old surveys for a final time in 1974. The redesigned surveys encountered surprisingly few operational problems in their initial year in spite of the fact that the old and redesigned surveys were carried out simultaneously. The field procedures and to a

lesser extent the data processing system for the three surveys have undergone a fine tuning in 1975 and 1976. The forecasts which were produced in these two years have proved to be very useful to the various marketing agencies. Their usefulness has been enhanced by their timely release to users following the completion of field work.

Looking ahead to future surveys, the introduction of a sample EA rotation scheme will be a major objective. Also, the importance of initiating and maintaining an adequate research program, which would continue to study the relationship between fruit counts and fruit weights as well as investigate other problems mentioned in the Introduction, cannot be overemphasized.

The author wishes to thank Dr. M.P. Singh and the referee for their helpful comments.

RESUME

Tous les ans depuis 1964, on a effectué des enquêtes sur les rendements prévus dans la péninsule du Niagara. On cherche ainsi à prévoir les variations de la production commercialisable dans la région par rapport à l'année précédente. Les estimations sont faites assez longtemps avant la récolte pour constituer des facteurs importants de négociation des prix entre les producteurs et les conditionneurs, ainsi que des indicateurs qui avertissent les organismes de vente de situations spéciales qui exigeraient une refonte immédiate des projets de vente. En 1973, on a entrepris de revoir en détail ces enquêtes. Le présent rapport comprend un sommaire du plan d'échantillonnage, des procédés de collecte des données et des procédés d'estimation qui ont été inclus dans les enquêtes révisées sur les rendements prévus des cerises aigres, des pêches et des raisins.

REFERENCES

[1] Cochran, W.G., (1963), "Sampling Techniques", John Wiley and Sons, Inc., New York, Second Edition, pp. 12-15.

[2] Hansen, M.H., Hurwitz, W.N., and Madow, W.G., (1953), "Sample Survey Methods and Theory", John Wiley and Sons, Inc., New York, Volume I.

[3] Jessen, R.J., (1955), "Determining the Fruit Count on a Tree by Randomized Branch Sampling", Biometrics, Volume 11, pp. 99-109.

[4] Murthy, M.N., (1967), "Sampling Theory and Methods", Statistical Publishing Society, Calcutta, pp. 376-378.

[5] Ontario Ministry of Agriculture and Food report, "1971 Fruit Tree Census, Part I - Grapes".

[6] Ontario Ministry of Agriculture and Food report, "1971 Fruit Tree Census, Part II - Tender Fruits".

APPENDIX A: Land Use Changes in the Niagara District

Table 4 illustrates the dynamic nature of the fruit tree and grape vine populations by providing an indication of the fruit tree and grape vine turnover rate in the Niagara district.

TABLE 4

Fruit Type	Age Range in Years	Age Group Total as a Percentage of Total Trees or Vines in 1971*
Grapes	1-3	12.08
	4-10	18.52
	11-20	34.95
	21&over	34.45
Peaches	1-3	28.48
	4-9	34.59
	10&over	36.93
Sour Cherries.	1-5	21.58
	6-10	14.84
	11-15	22.77
	16&over	40.81
Pears	1-10	26.15
	11-20	31.76
	21&over	42.09

SOURCE: Ontario Ministry of Agriculture and Food report, see [5] & [6].

*As in Table 3, these percentages were not calculated using 100% of the trees and vines in the Niagara district.

This table shows that peach trees are a particularly volatile population with 28.48% of the 1971 reported peach tree total being trees which were 1 to 3 years in age. This high percentage of young peach trees can be attributed to a number of factors, two of these being:

(i) The high incidence in the region at this time of such peach tree diseases as peach canker, which shortened the life expectancy of peach trees in relation to other fruit trees.

(ii) Frequent sizeable fluctuations in the prices paid by processors for peaches resulted in corresponding fluctuations in the total acreage in the region devoted to growing peaches.

It is also worthy of note that if a constant planting rate per year for peach trees is assumed then 9.5%, or over 70,000 peach trees, would have been planted in each of the years 1968 to 1970 inclusive.

APPENDIX B: Formulae Used to Estimate the Marketable Production Change and Measures of its Precision

The ratio of change estimator employed in each of the three objective yield surveys to estimate the marketable production change in the region from the previous year, denoted $\hat{R}$, is given by,

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \tag{1}$$

where: $\hat{Y}$ = the estimated total number of pieces of marketable fruit (i.e. sour cherries, peaches or grapes) in the region in the current year,

and $\hat{X}$ = the estimated total number of pieces of marketable fruit in the region in the previous year.

This ratio of change estimator is termed the combined ratio estimator since it is based on the pooled or combined estimators $\hat{Y}$ and $\hat{X}$. Now, the estimators $\hat{Y}$ and $\hat{X}$ are defined as,

$$\hat{Y} = \sum_{h=1}^{7} \hat{Y}_h = \sum_{h=1}^{7} \frac{1}{\omega_h} \sum_{r=1}^{\omega_h} \hat{Y}_{hr} \tag{2}$$

and

$$\hat{X} = \sum_{h=1}^{7} \hat{X}_h = \sum_{h=1}^{7} \frac{1}{\omega_h} \sum_{r=1}^{\omega_h} \hat{X}_{hr} \tag{3}$$

where: $\hat{Y}_h$ = the estimated total number of pieces of marketable fruit in stratum h in the current year,

$\hat{X}_h$ = identical to $\hat{Y}_h$ except that it refers to the previous year,

$\hat{Y}_{hr}$ = the estimated total number of pieces of marketable fruit in replicate r of stratum h in the current year,

$\hat{X}_{hr}$ = identical to $\hat{Y}_{hr}$ except that it refers to the previous year,

and ω_h = the number of replicates in stratum h.

Sour Cherry Survey

$$\bar{y}_{hrmijk} = \frac{1}{4} \sum_{\ell=1}^{4} \frac{y_\ell}{P_\ell} \tag{5}$$

where: y_ℓ = the total number of marketable sour cherries counted on the sample limb(s) selected on sample tree ℓ,

and P_ℓ = the probability of choosing the sample limb(s) selected on sample tree ℓ.

Peach Survey

$$\bar{y}_{hrmijk} = \frac{1}{4} \sum_{\ell=1}^{4} y_\ell \tag{6}$$

where: y_ℓ = the total number of marketable peaches counted on sample tree ℓ.

Grape Survey

$$\bar{y}_{hrmijk} = \frac{1}{5} \sum_{\ell=1}^{5} \{(g_\ell)(\frac{y_\ell}{5})\} \tag{7}$$

where: g_ℓ = the total number of bunches of grapes counted on sample vine ℓ,

and y_ℓ = the total number of marketable grapes counted in the 5 bunches selected from sample vine ℓ.

The formula for the corresponding replicate level estimator for the previous year, $\hat{X}_{hr}$, is identical to the equation (4) formula for $\hat{Y}_{hr}$, except that A_{hrmijk} and $\bar{y}_{hrmijk}$ are replaced by B_{hrmijk} and $\bar{x}_{hrmijk}$ respectively, which pertain to the previous year.

For the non-agricultural EA stratum (i.e. stratum 7), $\hat{Y}_{hr}$ is given by,

$$\hat{Y}_{hr} = \frac{N_h}{n_{hr}} \sum_{m=1}^{n_{hr}} (A^*_{rm} \cdot \bar{y}^*_{rm}) \tag{8}$$

For each of the agricultural EA strata (i.e. strata 1 to 6 inclusive), $\hat{Y}_{hr}$ is given by,

$$\hat{Y}_{hr} = \frac{N_h}{n_{hr}} \sum_{m=1}^{n_{hr}} \frac{T_{hrm}}{t_{hrm}} \sum_{i=1}^{t_{hrm}} \frac{Q_{hrmi}}{q_{hrmi}} \sum_{j=1}^{q_{hrmi}} \left\{ \frac{M_{hrmij}}{M_{hrmijk}} (A_{hrmijk} \cdot \bar{y}_{hrmijk}) \right\} \qquad (4)$$

where: N_h = the total number of EA's in the population in stratum h,

n_{hr} = the number of EA's in the sample in replicate r of stratum h,

T_{hrm} = the total number of first stage segments in EA m, in replicate r of stratum h,

t_{hrm} = the number of first stage segments selected in the sample from EA m, in replicate r of stratum h,

Q_{hrmi} = the total number of second stage segments in the first stage segment i of EA m, in replicate r of stratum h,

q_{hrmi} = the number of second stage segments selected in the sample from first stage segment i of EA m, in replicate r of stratum h,

M_{hrmij} = the total number of sour cherry or peach trees, or vine spaces in second stage segment j, in first stage segment i of EA m, in replicate r of stratum h, when this second stage segment was ground mapped,

M_{hrmijk} = the total number of sour cherry or peach trees, or vine spaces in selected block k of second stage segment j, in first stage segment i of EA m, in replicate r of stratum h, when this second stage segment was ground mapped,

A_{hrmijk} = the current year's count of the total number of producing sour cherry or peach trees, or vine spaces in selected block k of second stage segment j, in first stage segment i of EA m, in replicate r of stratum h,

and $\bar{y}_{hrmijk}$ = the current year's average number of pieces of marketable fruit per sample tree, or sample vine, for the sample of 4 producing trees, or 5 producing vines, chosen from selected block k of second stage segment j, in first stage segment i of EA m, in replicate r of stratum h.

NOTE: The variable $\bar{y}_{hrmijk}$ is defined in a different manner for each survey. The formula representation of this variable for each of the three surveys is as follows:

where: A^*_{rm} = the current year's count of the total number of producing sour cherry or peach trees, or vine spaces in EA m, in replicate r of stratum 7,

and $\bar{y}^*_{rm}$ = the current year's average number of pieces of marketable fruit per sample tree, or sample vine, for the sample of 4 producing trees, or 5 producing vines, chosen from EA m, in replicate r of stratum 7.

NOTE: Equations (5), (6) and (7) also define the variable $\bar{y}^*_{rm}$ for the three surveys.

For stratum 7, the formula for $\hat{X}_{hr}$ is identical to the equation (8) formula for $\hat{Y}_{hr}$, except that A^*_{rm} and $\bar{y}^*_{rm}$ are replaced by B^*_{rm} and $\bar{x}^*_{rm}$ respectively, which refer to the previous year.

Combining equations (4) and (8), equation (2) may be rewritten as,

$$\hat{Y} = \sum_{h=1}^{6} \hat{Y}_h + \hat{Y}_7$$

$$= \sum_{h=1}^{6} \frac{1}{\omega_h} \sum_{r=1}^{\omega_h} \frac{N_h}{n_{hr}} \sum_{m=1}^{n_{hr}} \frac{T_{hrm}}{t_{hrm}} \sum_{i=1}^{t_{hrm}} \frac{Q_{hrmi}}{q_{hrmi}} \tag{9}$$

$$\sum_{j=1}^{q_{hrmi}} \left\{ \frac{M_{hrmij}}{M_{hrmijk}} \left(A_{hrmijk} \cdot \bar{y}_{hrmijk} \right) \right\} + \frac{1}{\omega_h} \sum_{r=1}^{\omega_h} \frac{N_h}{n_{hr}} \sum_{m=1}^{n_{hr}} \left(A^*_{rm} \cdot \bar{y}^*_{rm} \right).$$

A similar expression could be written for $\hat{X}$.

As outlined by Murthy [4], an estimate of the variance of this ratio of change estimator, denoted $v(\hat{R})$, is given by,

$$v(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h=1}^{7} \frac{1}{\omega_h(\omega_h - 1)} \left\{ \sum_{r=1}^{\omega_h} (\hat{Y}_{hr} - \hat{R}\,\hat{X}_{hr})^2 - \omega_h\,(\hat{Y}_h - \hat{R}\,\hat{X}_h)^2 \right\}. \tag{10}$$

A measure of the absolute precision of this ratio of change estimator is provided by the standard error of the estimate. An estimate of the standard error of $\hat{R}$, denoted $s(\hat{R})$, is

$$s(\hat{R}) = \sqrt{v(\hat{R})}. \tag{11}$$

A measure of the relative precision of $\hat{R}$ is offered by the coefficient of variation of $\hat{R}$. An estimate of the coefficient of variation of $\hat{R}$, denoted $cv(\hat{R})$, expressed as a percentage is,

$$cv(\hat{R}) = \frac{s(\hat{R})}{\hat{R}} . 100\%. \tag{12}$$

The ratio of change estimator employed in each of the three surveys is, as are ratio estimators in general, a biased estimator. An estimate of the bias of this ratio of change estimator, denoted $b(\hat{R})$, is given by,

$$b(\hat{R}) = \frac{1}{\hat{X}^2} \sum_{h=1}^{7} \frac{1}{\omega_h(\omega_h - 1)} \left\{ \sum_{r=1}^{\omega_h} \left[\hat{X}_{hr}(\hat{R}\hat{X}_{hr} - \hat{Y}_{hr})\right] - \omega_h \hat{X}_h(\hat{R}\hat{X}_h - \hat{Y}_h)\right\} . \tag{13}$$

Unfortunately, no further discussion of the bias is possible at this time as a project designed to measure the bias and examine its effect on the accuracy of the ratio of change estimator has not yet been completed.

AN INVESTIGATION OF THE PROPERTIES OF RAKING RATIO ESTIMATORS: I WITH SIMPLE RANDOM SAMPLING

H.R. Arora and G.J. Brackstone
Census Survey Methods Division

The 1971 and 1976 Censuses of Population and Housing have utilized the raking ratio estimation procedure to obtain estimates for variables collected only on a sample basis. This paper derives large sample approximations for the bias and variance of such estimates and examines their performance in an empirical study.

1. INTRODUCTION

Both the 1971 and 1976 Censuses of Population and Housing in Canada utilized sampling. Within each Enumeration Area (EA) a systematic 1 in 3 sample of households was selected. Each sample household received a long questionnaire which contained both the basic (or 100%) questions and a set of additional sample questions. The remaining households received a short questionnaire that contained only the basic questions. The Raking Ratio Estimation Procedure (RREP) with four iterations was used to make estimates from the sample data. The purpose of this paper is to obtain asymptotic expressions for the bias and variance of the Raking Ratio Estimator for different numbers of iterations ($t=1,2,3,4$) and to present some empirical results describing the efficiency and bias of these estimators compared to a simple weighted estimator ($t=0$) when applied to the 1974 Test Census data.

2. THE RAKING RATIO ESTIMATOR OF A DOMAIN TOTAL

Suppose both the sample and population units are cross-classified in a two-dimensional matrix that is defined in terms of basic characteristics. Let n_{ij}, N_{ij} be respectively the sample and population counts in the (i,j)th cell of this matrix ($N = \sum_i \sum_j N_{ij}$, $n = \sum_i \sum_j n_{ij}$). Let y_{ijk} be the observed value of a certain sample characteristic, y, for the kth sample unit in the (i,j)th cell. Note that the variable y_{ijk} could be used to denote that the (i,j,k)th sample unit is in any specified category defined in terms of census variables. The following algebra, therefore, applies to any census cell value estimated from the census sample whether

the cell is in a tabulation of just one variable or in a cross-tabulation of several variables, and whether in a tabulation for all Canada or any smaller domain. The objective is to estimate the population total, Y, for the sample characteristic, y.

The estimator $\hat{Y}$ of Y will be of the form

$$\hat{Y} = \sum_i \sum_j W_{ij}\, y_{ij} \tag{2.1}$$

where $y_{ij} = \sum_k y_{ijk}$ and $W_{ij} = f\,(\{n_{ij}, N_{ij}\})$ is a weight attached to each sample unit in the (i,j)th cell.

Let $W_{ij}^{(t)}$ denote the weight, W_{ij}, corresponding to the t th iteration of the RREP. Then $W_{ij}^{(t)}$ is given by

$$W_{ij}^{(0)} = \frac{N}{n} \quad \text{for all } i,j$$

$$\left.\begin{aligned} W_{ij}^{(t)} &= W_{ij}^{(t-1)} \frac{N_{.j}}{\sum_a n_{aj}\, W_{aj}^{(t-1)}} && \text{if } t \text{ is even} \\ &\; W_{ij}^{(t-1)} \frac{N_{i.}}{\sum_b n_{ib}\, W_{ib}^{(t-1)}} && \text{if } t \text{ is odd.} \end{aligned}\right. \tag{2.2}$$

where a subscript dot (.) indicates summation over the replaced subscript. If t is even the weighted sample total in each column of the matrix is exactly equal to the known population total for the column, while, if t is odd, this equality is exact for the rows of the matrix.

$$\text{Let} \quad \hat{Y}^{(t)} = \sum_i \sum_j W_{ij}^{(t)}\, y_{ij}. \tag{2.3}$$

Before going on to derivations of the bias and variance of $\hat{Y}^{(t)}$ for various values of t we will define some notation that will shorten many of the subsequent expressions.

Let Y_{ij} = population total for variable Y in the (i,j)th cell

$$R_{i.} = Y_{i.}/N_{i.}\ , \quad R_{.j} = Y_{.j}/N_{.j}\ , \quad \rho_{ij} = N_{ij}/N_{i.}\ , \quad \kappa_{ij} = N_{ij}/N_{.j}\ ,$$

$$\gamma_{ia} = \sum_j \kappa_{ij}\,\rho_{aj}\ , \quad \alpha_i = N_{i.}^2\left(\frac{1}{n_{i.}} - \frac{1}{N_{i.}}\right)\ , \quad \hat{N}_{ij}^{(t)} = n_{ij}W_{ij}^{(t)}\ , \quad \hat{Y}_{ij}^{(t)} = y_{ij}W_{ij}^{(t)}\ ,$$

$$N_{ij}^{(t)} = E(\hat{N}_{ij}^{(t)})\ , \quad Y_{ij}^{(t)} = E(\hat{Y}_{ij}^{(t)})\ , \quad \hat{R}_{i.}^{(t)} = \hat{Y}_{i.}^{(t)}/\hat{N}_{i.}^{(t)}\ , \quad \hat{R}_{.j}^{(t)} = \hat{Y}_{.j}^{(t)}/\hat{N}_{.j}^{(t)}\ ,$$

$$R_{i.}^{(t)} = E(\hat{Y}_{i.}^{(t)})/E(\hat{N}_{i.}^{(t)})\ , \quad R_{.j}^{(t)} = E(\hat{Y}_{.j}^{(t)})/E(\hat{N}_{.j}^{(t)})\ , \quad \hat{\rho}_{ij}^{(t)} = \hat{N}_{ij}^{(t)}/\hat{N}_{i.}^{(t)}\ ,$$

$$\hat{\kappa}_{ij}^{(t)} = \hat{N}_{ij}^{(t)}/\hat{N}_{.j}^{(t)}\ , \quad \rho_{ij}^{(t)} = E(\hat{N}_{ij}^{(t)})/E(\hat{N}_{i.}^{(t)})\ , \quad \kappa_{ij}^{(t)} = E(\hat{N}_{ij}^{(t)})/E(\hat{N}_{.j}^{(t)})\ ,$$

where i and a have the same range.

For any variables V,W taking the values V_{iu}, W_{iu}, for the u th unit in the ith row we shall define

$$S_i^2\,(V) = \frac{1}{N_{i.} - 1} \sum_u (V_{iu} - \bar{V}_i)^2$$

$$S_i\,(V,W) = \frac{1}{N_{i.} - 1} \sum_u (V_{iu} - \bar{V}_i)(W_{iu} - \bar{W}_i) \qquad \text{for the population,}$$

and

$$s_i^2\,(v) = \frac{1}{n_{i.} - 1} \sum_u (v_{iu} - \bar{v}_i)^2$$

$$s_i(v,w) = \frac{1}{n_{i.} - 1} \sum_u (v_{iu} - \bar{v})(w_{iu} - \bar{w}_i) \qquad \text{for the sample,}$$

where $\bar{V}_i$, $\bar{v}_i$ are respectively the population and sample means of V in the ith row.

From (3.2) $\quad \hat{\rho}_{ij}^{(o)} = \hat{\rho}_{ij}^{(1)} = n_{ij}/n_{i.}$ and will be denoted by $\hat{\rho}_{ij}$.

Also $\quad \hat{\kappa}_{ij}^{(1)} = \hat{N}_{ij}^{(1)}/\hat{N}_{.j}^{(1)}$ and will be denoted by $\hat{\kappa}_{ij}$.

We define $\quad \hat{\gamma}_{ia} = \sum_j \hat{\kappa}_{ij}\,\hat{\rho}_{aj}$.

It should be noted that $R_{i.}^{(1)} = R_{i.}$, $R_{.j}^{(1)} = R_{.j}$, $\rho_{ij}^{(1)} = \rho_{ij}$, and $\kappa_{ij}^{(1)} = \kappa_{ij}$.

3. ASSUMPTIONS ABOUT SAMPLE SELECTION

The RREP is carried out separately at a Weighting Area (WA) level where each WA consists of a set of whole EA's - usually 4-8. The Census sampling scheme within a WA is therefore a stratified systematic sample with respect to household characteristics, and a stratified systematic sample of clusters with respect to person characteristics. In this paper we shall assume simple random sampling without replacement within the WA. Because the effect of clustering for person characteristics is regarded as significant, these results are therefore applicable only to characteristics defined at the household level. Neither the stratification effect (EA's within WA's) nor the systematic sampling effect (within EA's) are considered here. Although the effect of systematic sampling is probably negligible further research would be required to assess the effect of stratification. The corresponding formulae for cluster sampling have been derived and will be presented with empirical data in a separate paper.

4. BIAS AND VARIANCE OF THE NO-ITERATION ESTIMATOR

$$\hat{Y}^{(0)} = \sum_i \sum_j \frac{N}{n} y_{ij} = N\bar{y}, \quad E(\hat{Y}^{(0)}) = Y, \tag{4.1}$$

so that $\hat{Y}^{(0)}$ is an unbiased estimator of Y. Now

$$V(\hat{Y}^{(0)}) = N^2(\frac{1}{n} - \frac{1}{N}) S^2 (y) \tag{4.2}$$

where
$$S^2(y) = \frac{1}{N-1} \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y})^2.$$

An unbiased estimator of $V(\hat{Y}^{(o)})$ is given by

$$v(\hat{Y}^{(0)}) = N^2 (\frac{1}{n} - \frac{1}{N}) s^2(y) \tag{4.3}$$

where
$$s^2(y) = \frac{1}{n-1} \sum_i \sum_j \sum_k (y_{ijk} - \bar{y})^2, \quad (n \geq 2). \tag{4.3}$$

5. BIAS AND VARIANCE OF THE ONE-ITERATION ESTIMATOR

$$\hat{Y}^{(1)} = \sum_{ij}\sum \frac{N_{i.}}{n_{i.}} y_{ij} = \sum_i N_{i.}\ \bar{y}_{i.}. \tag{5.1}$$

so that, noting for a given $n_{i.}$, we have SRS from each row, we get

$$E(\hat{Y}^{(1)}) = E_1 \sum_i N_{i.} E_2\ (\bar{y}_{i.}) = E_1 (\sum_i N_{i.} \bar{Y}_{i.}) = E_1(Y) = Y,$$

where a subscript 2 will denote expectations conditioned on the set of $n_{i.}$, while the subscript 1 will denote expectations over $n_{i.}$.

Note that the formation of the cross-classification matrix ensure that $n_{i.} > 0.$.

Now $V(\hat{Y}^{(1)}) = E_1 V_2(\hat{Y}^{(1)})$ since $V_1 E_2(\hat{Y}^{(1)}) = 0$. Therefore,

$$V(\hat{Y}^{(1)}) = \sum_i S_i^2\ (y)\ E_1(\alpha_i). \tag{5.2}$$

An unbiased estimator of $V(\hat{Y}^{(1)})$ is given by

$$v(\hat{Y}^{(1)}) = \sum_i \alpha_i\ s_i^2(y), \quad (n_{i.} \geq 2). \tag{5.3}$$

6. BIAS AND VARIANCE OF THE TWO-ITERATION ESTIMATOR

In this section, we derive the bias and variance results previously stated in [2].

$$\hat{Y}^{(2)} = \sum_i \sum_j N_{.j} \frac{\hat{Y}_{ij}^{(1)}}{\hat{N}_{.j}^{(1)}} = \sum_j N_{.j} \frac{\hat{Y}_{.j}^{(1)}}{\hat{N}_{.j}^{(1)}}. \tag{6.1}$$

6.1 Asymptotic Bias of $Y^{(2)}$

Using the large sample results for the expected value of a ratio estimator we obtain

$$B(\hat{Y}^{(2)}) \doteq -\sum_j \frac{1}{N_{.j}} \text{Cov}\ (\hat{N}^{(1)}_{.j},\ Y^{(1)}_{.j} - R^{(1)}_{.j}\ \hat{N}^{(1)}_{.j})$$

$$= -\sum_j \frac{1}{N_{.j}}\, C(\sum_i \frac{N_{i.}}{n_{i.}} \sum_{u=1}^{n_{i.}} {}_j a_{iu},\ \sum_i \frac{N_{i.}}{n_{i.}} \sum_{u=1}^{n_{i.}} {}_j a_{iu}\ (z_{iu} - R^{(1)}_{.j})) \qquad (6.2)$$

where $\quad {}_j a_{iu} \quad = \quad$ 1 if the uth unit in the ith row belongs to the jth column
0 otherwise

and $\quad z_{iu} \quad = \quad$ value of y for the uth unit in the ith row.

Since $E_2(\hat{N}^{(1)}_{.j})$ is a constant and is therefore uncorrelated with $E_2(\hat{Y}^{(1)}_{.j} - R^{(1)}_{.j}\ \hat{N}^{(1)}_{.j})$, the covariance on the right hand side of (6.2) is equal to $E_1 C_2\ (\hat{N}^{(1)}_{.j},\ \hat{Y}^{(1)}_{.j} - R^{(1)}_{.j}\hat{N}^{(1)}_{.j})$. Therefore

$$B(\hat{Y}^{(2)}) \quad \doteq \quad -\sum_j \frac{1}{N_{.j}}\ E_1[\sum_i \alpha_i\ S_i\ (\ {}_j a,\ {}_j a\ (z - R^{(1)}_{.j}))] \qquad (6.3)$$

$$= \quad -\sum_j \frac{1}{N_{.j}}\, E_1 \sum_i \frac{\alpha_i N_{i.}}{N_{i.} - 1}\ (1 - \rho_{ij})(\frac{Y_{ij}}{N_{i.}} - \rho_{ij}\ R^{(1)}_{.j}) \qquad (6.4)$$

after some algebra.

An estimator of $B(Y^{(2)})$ can be obtained by substituting the sample covariance $s_i({}_j a,\ {}_j a\ (z - \hat{R}^{(1)}_{.j}))$ in (6.3) giving

$$b(\hat{Y}^{(2)}) \quad = \quad -\sum_j \frac{1}{\hat{N}^{(1)}_{.j}}\ \sum_i \alpha_i\ \frac{n_{i.}}{n_{i.} - 1}\ (1 - \hat{\rho}_{ij})(\frac{y_{ij}}{n_{i.}} - \hat{\rho}_{ij}\ \hat{R}^{(1)}_{.j}) \qquad (6.5)$$

6.2 Asymptotic Variance of $Y^{(2)}$

For large sample size n,

$$V(y^{(2)}) \doteq V\Big[\sum_j (Y^{(1)}_{.j} - R^{(1)}_{.j}\ N^{(1)}_{.j})\Big]$$

$$\doteq E_1 V_2 \Big[\sum_j (\hat{Y}^{(1)}_{.j} - R^{(1)}_{.j}\ \hat{N}^{(1)}_{.j})\Big]$$

$$= E_1 V_2 \Big(\sum_i \frac{N_{i.}}{n_{i.}} \sum_{u=1}^{n_{i.}} [z_{iu} - \sum_j R^{(1)}_{.j}\ {}_j a_{iu}]\Big)$$

$$= E_1 \sum_i \alpha_i S_i^2 (z - \sum_j {}_j a\ R^{(1)}_{.j}), \tag{6.6}$$

from which we obtain

$$V(\hat{Y}^{(2)}) = V(\hat{Y}^{(1)}) + E_1 \sum_i \alpha_i \frac{N_{i.}}{N_{i.}-1} [\sum_j \rho_{ij}\ R^{(1)^2}_{.j}$$

$$- \frac{2}{N_{i.}} \sum_j R^{(1)}_{.j} (Y_{ij} - \rho_{ij}\ Y_{i.}) - (\sum_j \rho_{ij}\ R^{(1)}_{.j})^2]. \tag{6.7}$$

An estimator of $V(\hat{Y}^{(2)})$ is given by

$$v(\hat{Y}^{(2)}) = \sum_i \alpha_i\ s_i^2 (z - \sum_j {}_j a\ \hat{R}^{(1)}_{.j})$$

$$= v(\hat{Y}^{(1)}) + \sum_i \alpha_i \frac{n_{i.}}{n_{i.}-1} [\sum_j \hat{\rho}_{ij}\ \hat{R}^{(1)^2}_{.j}$$

$$- \frac{2}{n_{i.}} \sum_j \hat{R}^{(1)}_{.j} (y_{ij} - \hat{\rho}_{ij}\ y_{i.}) - (\sum_j \hat{\rho}_{ij}\ \hat{R}^{(1)}_{.j})^2]. \tag{6.8}$$

7. VARIANCE OF THE THREE AND FOUR-ITERATION ESTIMATORS

$$\hat{Y}^{(3)} = \sum_i \sum_j N_{i.} \frac{\hat{Y}^{(2)}_{ij}}{\hat{N}^{(2)}_{i.}} = \sum_i N_{i.} \frac{\hat{Y}^{(2)}_{i.}}{\hat{N}^{(2)}_{i.}}\ . \tag{7.1}$$

7.1 Asymptotic Variance of $\hat{Y}^{(3)}$

For large sample size n,

$$V(\hat{Y}^{(3)}) \doteq E_1 V_2\Big(\sum_i (\hat{Y}_{i.}^{(2)} - R_{i.}^{(2)} \hat{N}_{i.}^{(2)})\Big)$$

$$\doteq E_1 V_2 [\sum_i \frac{N_{i.}}{n_{i.}} \sum_{u=1}^{n_{i.}} (z_{iu} - \sum_j R_{.j}^{(1)} {}_j a_{iu}) - \sum_i N_{i.} R_{i.}^{(2)}$$

$$+ \sum_i \frac{N_{i.}}{n_{i.}} \sum_{u=1}^{n_{i.}} \sum_j {}_j a_{iu} \sum_a R_{a.}^{(2)} \kappa_{aj}]$$

$$= E_1 [\sum_i \alpha_i S_i^2 (z - \sum_j {}_j a(R_{.j}^{(1)} - \sum_a R_{a.}^{(2)} \kappa_{aj}))], \qquad (7.2)$$

from which we obtain

$$V(\hat{Y}^{(3)}) \doteq V(\hat{Y}^{(2)}) + E_1 [\sum_i \alpha_i \frac{N_{i.}}{N_{i.}-1} [\sum_j \rho_{ij} (\sum_a R_{a.}^{(2)} \kappa_{aj})^2 \qquad (7.3)$$

$$- (\sum_a R_{a.}^{(2)} \gamma_{ai})^2 + 2 \sum_j (\frac{Y_{ij}}{N_{i.}} - \rho_{ij} R_{.j}^{(1)})(\sum_a R_{a.}^{(2)} \kappa_{aj})$$

$$- 2(R_{i.} - \sum_j \rho_{ij} R_{.j}^{(1)})(\sum_a R_{a.}^{(2)} \gamma_{ai})]].$$

An estimator of $V(\hat{Y}^{(3)})$ obtained by substituting the sample variance in (7.2) is given by

$$v(\hat{Y}^{(3)}) = v(\hat{Y}^{(2)}) + \sum_i \alpha_i \frac{n_{i.}}{n_{i.}-1} [\sum_j \hat{\rho}_{ij} (\sum_a \hat{R}_{a.}^{(2)} \hat{\kappa}_{aj})^2 - (\sum_a \hat{R}_{a.}^{(2)} \hat{\gamma}_{ai})^2$$

$$+ 2 \sum_j (\frac{y_{ij}}{n_{i.}} - \hat{\rho}_{ij} \hat{R}_{.j}^{(1)})(\sum_a \hat{R}_{a.}^{(2)} \hat{\kappa}_{aj})$$

$$- 2 (\frac{y_{i.}}{n_{i.}} - \sum_j \hat{\rho}_{ij} \hat{R}_{.j}^{(1)})(\sum_a \hat{R}_{a.}^{(2)} \hat{\gamma}_{ai})], \qquad (7.4)$$

By the procedure of repeatedly linearizing ratios similar to that used above we obtain the following results

$$V(\hat{Y}^{(4)}) = V(\hat{Y}^{(3)}) + E_1[\sum_i \alpha_i \frac{N_{i.}}{N_{i.}-1} \{\sum_j \rho_{ij} (\sum_a \sum_b \kappa_{aj} \rho_{ab}^{(2)} R_{.b}^{(3)})^2$$

$$- (\sum_a \sum_b R_{.b}^{(3)} \rho_{ab}^{(2)} \gamma_{ai})^2$$

$$- 2 \sum_j (\frac{Y_{ij}}{N_{i.}} - \rho_{ij} (R_{.j}^{(1)} - \sum_a R_{a.}^{(2)} \kappa_{aj})) (\sum_a \sum_b \kappa_{aj} \rho_{ab}^{(2)} R_{.b}^{(3)})$$

$$+ 2(R_{i.} - \sum_j \rho_{ij} (R_{.j}^{(1)} - \sum_a R_{a.}^{(2)} \kappa_{aj}))(\sum_a \sum_b R_{.b}^{(3)} \rho_{ab}^{(2)} \gamma_{ai})\}]. \qquad (7.5)$$

An estimator of $V(Y^{(4)})$ obtained by substituting the sample variance in (7.5) is given by

$$v(\hat{Y}^{(4)}) = v(\hat{Y}^{(3)}) + \sum_i \alpha_i \frac{n_{i.}}{n_{i.}-1} \{\sum_j \hat{\rho}_{ij} (\sum_a \sum_b \hat{R}_{.b}^{(3)} \hat{\kappa}_{aj} \hat{\rho}_{ab}^{(2)})^2$$

$$- (\sum_a \sum_b \hat{R}_{.b}^{(3)} \hat{\rho}_{ab}^{(2)} \hat{\gamma}_{ai})^2$$

$$- 2 \sum_j (\frac{Y_{ij}}{n_{i.}} - \hat{R}_{.j}^{(1)} \hat{\rho}_{ij} + \hat{\rho}_{ij} \sum_a \hat{R}_{a.}^{(2)} \hat{\kappa}_{aj}) (\sum_a \sum_b \hat{\kappa}_{aj} \hat{\rho}_{ab}^{(2)} \hat{R}_{.b}^{(3)})$$

$$+ 2 (\frac{y_{i.}}{n_{i.}} - \sum_j \hat{\rho}_{ij} \hat{R}_{.j}^{(1)} + \sum_a \hat{R}_{a.}^{(2)} \hat{\gamma}_{ai}) (\sum_a \sum_b \hat{R}_{.b}^{(3)} \hat{\gamma}_{ai} \hat{\rho}_{ab}^{(2)})\}. \qquad (7.6)$$

Similar but longer expressions for $B(\hat{Y}^{(3)})$ and $B(\hat{Y}^{(4)})$ and their estimators have also been derived.

8. THE EMPIRICAL INVESTIGATION

To compare the variance and bias of raking ratio estimators for various iterations and for various characteristics, the formulae derived in the previous sections were applied to data from the 1974 Canadian Test Census. This Test Census utilized the sampling procedure described in section 1 of this paper. The data used in the empirical investigations came from one Electoral District (ED) which contained 15 Weighting Areas (WA's). The WA's varied in size from a low of 628 households (213 in the sample) to a high of 1946 households (647 in the sample). The average size was 1262 households.

The initial cross-classification table or weighting matrix for households is given in Appendix 1. The rows and columns of the initial table were collapsed if the following conditions were not met:

1. all $N_{i.}$ and $N_{.j}$ ≥ 35
2. all ratios $N_{i.}/n_{i.}$ and $N_{.j}/n_{.j}$ >1 but < 12
3. all $n_{i.}$ and $n_{.j} > 0$.

A collapsing strategy pre-defined the rows or columns to be collapsed when these conditions were not met. The collapsing procedure continued until all the above conditions were satisfied or until all steps in the collapsing strategy had been exhausted. The result was that the final weighting matrices would in general differ slightly from WA to WA. The largest weighting matrix had 18 rows and 4 columns while the smallest had 7 rows and 3 columns (compared to 20 rows and 4 columns in the initial matrix). Differences in collapsing between WA's explain some of the differences in variance reduction in the results presented below.

The variables collected in the 1974 Test Census were the following:

100%	Sample (33-1/3%)
Relation to Head	School Attendance
Sex	Years of Schooling
Age	Post-Secondary Education
Marital Status	Academic Qualifications
Mother Tongue	Labour Force Status
Type of Dwelling	Address Five Years Ago
Tenure	

Since the theory developed here is appropriate only to variables defined for the household our studies were restricted to the two housing variables (type of dwelling and tenure) and to the personal characteristics of the head of the household as attributes of the household (e.g., households with heads not in the labour force). Since each category of each of the above variables, and each cell of each cross-tabulation is a potential characteristic (y) that could be examined (and all of these at any geographic level within the WA), some arbitrary selection of characteristics was essential. Altogether 27 categories were investigated and the results presented here represent a typical cross-section of these 27 categories.

Clearly our primary interest is in the sample variables. However, a selection of 100% variables has been considered for two reasons. First, in cross-tabulations of sample variables against 100% variables estimates for 100% variables are published (though the variance of such estimates may be of little interest when the 100% value, and therefore the exact sampling deviation is available). Secondly, 100% variables can be used to some extent as examples of variables with a relatively high correlation with the variables used to define the weighting matrix in order to see the effects of the raking ratio procedure for such variables. For the purposes of presentation we have broken down the categories considered into three classes.

A. Categories defined by sample variables.

B. Categories defined by 100% variables but not used for control in the weighting matrix.

C. Categories defined by 100% variables and used for control in the weighting matrix.

The specific categories for which results are presented are the following:

A1 Households with Employed Heads

A2 Households with Unemployed Heads

A3 Households with Heads Not in Labour Force

A4 Households with Heads Not Moved in 5 Years

A5 Households with Heads Moved in Last 5 Years in Same Municipality

A6 Highest Grade of Head is 1 to 10

A7 Heads with Bachelor Degree or Higher.

B1 Households with 3 or 4 Persons

B2 Age of Head is Less Than 25

B3 Age of Head is 25 to 34

B4 Heads Who are Widowed, Divorced, or Separated.

C1 Households with 2 or Fewer Persons

C2 Age of Head is 65 or More

C3 Owned Dwellings

C4 Rented Apartments.

In addition to the categories listed above, all of which are geographically at the WA level, we also examined certain categories at the EA level within WA 2. This WA contained three EA's. The number of households in each of these EA's was respectively 451, 506, and 159 (152, 168, and 50 in the sample). The categories considered at the EA level were A1, A4, A6, B1 and B3.

Note that the same weights, calculated at the WA level, are used in producing estimates at the EA level.

Tables A1 to C4 summarize the results. These tables list for each iteration (i = 0, 1, 2, 3, 4) estimates of the population totals ($\hat{Y}^{(i)}$), standard errors (SE_i), coefficients of variation (CV_i), and the "ratio of error" (RE_i) defined as the standard error of the i-iteration estimator expressed as a % of the standard error of the no-iteration estimator. These figures are given for a sample of 5 WA's arranged in size from largest to smallest. Corresponding figures are also given at the Electoral District level.

These act as a summary of the WA data. A measure of the change in population estimates between iterations is also given. It is defined as

$$C(k-p) = \sqrt{\frac{\sum_{m=1}^{15} (\hat{Y}_m^{(p)} - \hat{Y}_m^{(k)})^2}{\sum_{m=1}^{15} (\hat{Y}_m^{(k)})^2}} \times 100\%$$

where $\hat{Y}_m^{(p)}$ is the population estimate at the pth iteration for weighting area m. C(k-p) can be thought of as an average of the percentage change in the population estimates between the kth and pth iterations since, if $\left|\hat{Y}_m^{(p)} - \hat{Y}_m^{(k)}\right| = r\hat{Y}_m^{(k)}$ (m=1, ..., 15), then C(k-p) = r x 100%.

One final point about the empirical results has to be noted. The iterative raking procedure can begin with either the rows or columns of the weighting matrix (the theory in the earlier sections of this paper started the procedure with the rows, i.e. $\hat{N}_{i.}^{(1)} = N_{i.}$). The choice of start will usually depend on the relative importance of final agreement for the rows and for the columns of the matrix. In the 1974 Test Census the iterative procedure began with the columns (i.e. $\hat{N}_{.j}^{(1)} = N_{.j}$) and therefore after four iterations ended with the rows (i.e. $\hat{N}_{i.}^{(4)} = N_{i.}$) of the weighting matrix in Appendix 1.

Because of the high computational cost, the bias estimates were calculated for each iteration in three WA's of different sizes. For 15 categories considered it was found that the absolute value of the estimated bias as a percentage of the population estimate never exceeded 0.24% at WA level. At the EA level it was found that this ratio never exceeded 0.39%. These figures are small compared with the corresponding coefficient of variation figures. Thus bias can be regarded as negligible.

9. ANALYSIS OF RESULTS

The principal findings from an examination of the empirical results are listed below.

a) As mentioned in the previous section the biases of the raking-ratio estimates appear to be negligible compared with their standard errors.

b) For A-type categories most of the gains in efficiency at the ED

level occured at the second iteration. An examination of the variables used in the rows and columns of the cross-classification tables (Appendix 1) reveals that some correlation between many of the tables is not unexpected.

Beyond the second iteration there is little further gain in efficiency and little change in the values of the estimates.

c) Given that there is little gain in efficiency for A-type (i.e. sample) variables beyond the second iteration, the justification for proceeding to four iterations is primarily in terms of improving the sample-population agreement for B and C-type (i.e. 100%) variables.

d) A comparison of CV_o and RE_4 for the individual A-type categories reveals a strong tendency for large gains in efficiency (i.e. low RE_4) to be associated with large categories (i.e. small CV_o). The exceptions to this tendency (e.g. A_3) are generally explicable by a strong association between a small sample category and the rows defining the cross-classification table.

e) As expected, the gains in efficiency for B-type categories depend heavily on the relationship of the category to the rows and columns of the cross-classification table and on the frequency of collapsing of rows and columns. For example, consider category B_3 (Age of the Head is 25 to 34). The category accounts for 75% of Heads in Age Groups 15-34, while Heads in Age groups 15-34 in turn account for all Heads in a certain group of rows. Since no collapsing took place over age groups (except for two or more persons non-family households), the gains for this category are high. On the other hand consider the larger category B_1 (Households with 3 or 4 persons). This category accounts for less than one-third of the category Households with 3 or more persons, which in turn accounts for all

Households in a certain group of rows. However, the category 3 or more persons is frequently collapsed with category 2 or less persons with the result that the gains in efficiency are lower for B_1 then for B_3, even though B_1 is a larger category.

f) The gains in efficiency for C-type categories depend heavily upon the collapsing that took place within each WA. Where no collapsing of important rows or columns took place, variances at the WA level are clearly reduced to zero for some categories at certain iterations. The results for C-type categories highlight the need to choose carefully the collapsing criteria and strategy since these will have a profound effect on the variance, and on the sample - 100% agreement, for certain C-type categories.

g) Categories at the EA level (or any geographic level below the WA) can be regarded as small WA categories. As expected, the gains in efficiency at the EA level are found to be smaller than for the corresponding categories at the WA level.

10. CONCLUSION

The results of the empirical study have shown some significant reductions in variance for sample values through the use of the raking-ratio estimation procedure. For each specific variable the majority of the reduction in variance occurs at one particular iteration (the first or the second) so that an appropriate one-dimensional ratio estimation procedure (i.e. a post-stratification) could produce comparable reductions in variance for each variable separately. However, the requirement for consistency between different estimates from the Census sample dictates the use of a single estimation or weighting procedure for all variables. The two-dimensional raking ratio estimation procedure enables the potential gains of ratio estimation to be realised for many variables simultaneously.

As described in section 3 the most important extension of this paper is to the case of cluster sampling so that results for the individual person characteristics collected on the Census sample can be obtained. A second direction for research is to investigate the effects of using different WA's and different weighting matrices. For example, the relative efficiency of using smaller WA's and therefore a less detailed weighting matrix could be examined.

11. ACKNOWLEDGEMENT

The authors wish to acknowledge the valuable assistance of Ian MacMillan and Michael Bankier in the writing of computer programs and checking of results, and of Professor J.N.K. Rao for his advice and contributions to this study.

RESUME

Au cours de l'analyse des données des recensements de la population et du logement de 1971 et 1976, on a utilisé le procédé d'estimation de l'échantillon en formation pour effectuer des estimations mettant en jeu des données obtenues par échantillonnage. Dans cet article, on fait l'approximation du biais et de la variance de ces estimations à partir de grands échantillons et on examine l'efficacité de cet estimateur dans une étude empirique.

REFERENCES

[1] Ireland, C.T. and Kullback, S., (1968), "Contingency Tables With Given Marginals", Biometrika, 55, 179-188.

[2] Brackstone G.J. and Rao J.N.K., (1976). "Raking Ratio Estimators, Survey Methodology, Volume 2, Number 1, 63-69.

APPENDIX 1

TABLE HH1:
Cross-Classification Table For
Housing and Household Weights

Household Type	Sex of Head	Age of Head	Number of Persons	Row No.	OWNED: Single Detached Column 1	OWNED: Other Column 2	RENTED: Apartment Column 3	RENTED: Other Column 4
ONE OR MORE FAMILY HOUSEHOLDS	MALE	15-34	LE 2	1				
			GT 2	2				
		35-64	LE 2	3				
			GT 2	4				
		GE 65	LE 2	5				
			GT 2	6				
	FEMALE	15-34	LE 2	7				
			GT 2	8				
		35-64	LE 2	9				
			GT 2	10				
		GE 65	LE 2	11				
			GT 2	12				
ONE PERSON NON-FAMILY HOUSEHOLDS	MALE	15-34		13				
		35-64		14				
		GE 65		15				
	FEMALE	15-34		16				
		35-64		17				
		GE 65		18				
TWO-OR-MORE PERSON NON-FAMILY HOUSEHOLDS	MALE			19				
	FEMALE			20				

CATEGORIES AT WA LEVEL

A1

HOUSEHOLDS WITH EMPLOYED HEADS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
0.27	1.46	0.16	0.14	1.40	0.03

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	1,209.1	30.3	0.025	1,207.6	30.4	0.025	100.3	1,191.3	22.8	0.019	75.2	1,191.4	22.8	0.019	75.2	1,191.1	22.8	0.019	75.2
9	1,141.4	24.8	0.022	1,142.2	24.7	0.022	99.7	1,140.4	19.4	0.017	78.3	1,141.0	19.4	0.017	78.3	1,140.6	19.4	0.017	78.3
11	1,111.0	18.0	0.016	1,111.5	18.0	0.016	99.9	1,107.5	15.1	0.014	83.9	1,107.5	15.1	0.014	83.7	1,107.8	15.1	0.014	83.6
2	920.0	18.1	0.020	918.7	18.2	0.020	100.7	921.9	13.9	0.015	77.0	922.1	13.9	0.015	77.0	922.2	13.9	0.015	77.0
12	545.5	11.8	0.022	545.8	11.8	0.022	99.6	537.5	9.9	0.018	83.5	538.1	9.9	0.018	83.5	537.7	9.9	0.018	83.5
ED	13,873.7	84.0	0.006	13,878.3	83.6	0.006	99.5	13,844.3	65.7	0.005	78.2	13,852.6	65.8	0.005	78.3	13,845.9	65.7	0.005	78.2

A2

HOUSEHOLDS WITH UNEMPLOYED HEADS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.71	2.41	0.54	0.26	2.58	0.41

8	75.2	12.1	0.160	76.3	12.3	0.161	101.7	76.2	12.1	0.159	100.4	76.0	12.1	0.159	100.4	76.2	12.1	0.159	100.4
9	57.2	10.5	0.184	56.9	10.5	0.185	99.6	57.2	10.4	0.183	99.1	57.2	10.4	0.183	99.1	57.2	10.4	0.183	99.1
11	17.7	5.8	0.331	17.7	5.8	0.330	99.8	17.3	5.8	0.335	99.0	17.2	5.8	0.337	99.0	17.3	5.8	0.335	99.0
2	18.1	6.0	0.331	17.9	5.9	0.331	98.8	17.8	5.9	0.331	98.2	17.8	5.9	0.331	98.2	17.8	5.9	0.331	98.2
12	17.7	5.8	0.328	17.5	5.8	0.330	99.7	16.3	5.7	0.348	97.9	16.3	5.7	0.348	97.9	16.3	5.7	0.348	97.9
ED	542.0	32.4	0.060	543.4	32.6	0.060	100.7	543.4	32.3	0.059	99.7	541.7	32.3	0.060	99.7	542.5	32.3	0.059	99.7

A3

HOUSEHOLDS WITH HEADS NOT IN LABOUR FORCE

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
0.71	4.31	0.43	0.37	4.15	0.09

8	658.7	29.6	0.045	658.9	29.7	0.045	100.3	676.0	21.8	0.032	73.7	676.2	21.8	0.032	73.7	676.3	21.8	0.032	73.6
9	355.4	23.5	0.066	354.8	23.4	0.066	99.6	356.3	17.6	0.049	74.8	355.5	17.6	0.049	74.7	355.9	17.6	0.049	74.7
11	177.3	17.3	0.098	176.9	17.3	0.098	100.0	181.2	14.2	0.078	82.0	181.3	14.2	0.078	81.9	181.1	14.2	0.078	81.8
2	178.0	17.4	0.098	179.4	17.5	0.098	100.7	176.3	13.3	0.075	76.2	176.1	13.2	0.075	76.1	176.0	13.2	0.075	76.1
12	64.9	10.7	0.164	64.7	10.7	0.165	99.9	74.2	8.6	0.116	80.8	73.7	8.6	0.117	80.9	74.1	8.6	0.116	80.9
ED	4,502.5	80.9	0.018	4,496.4	80.6	0.018	99.6	4,531.2	61.4	0.014	75.8	4,524.2	61.5	0.014	75.9	4,530.3	61.3	0.014	75.8

A4

HOUSEHOLDS WITH HEADS NOT MOVED IN 5 YEARS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.63	1.90	0.62	0.27	1.51	0.37

8	890.3	31.2	0.035	879.5	26.5	0.030	84.9	884.6	24.9	0.028	80.1	890.1	24.4	0.027	78.2	887.1	24.4	0.027	78.2
9	804.1	28.0	0.035	802.5	24.4	0.030	87.2	809.6	23.1	0.029	82.5	806.3	22.7	0.028	81.2	807.9	22.7	0.028	81.1
11	650.1	25.3	0.039	652.6	23.8	0.036	94.0	674.9	21.8	0.032	86.4	670.8	21.8	0.033	86.2	672.2	21.8	0.032	86.1
2	304.0	21.2	0.069	305.3	19.7	0.065	93.2	304.7	18.9	0.062	89.5	306.1	18.8	0.061	88.8	305.6	18.6	0.061	88.8
12	241.8	17.1	0.071	245.2	14.2	0.058	83.3	252.0	13.7	0.054	80.5	248.9	13.6	0.054	79.5	249.4	13.6	0.054	79.5
ED	8,528.1	94.7	0.011	8,488.0	85.5	0.010	90.3	8,470.0	80.5	0.009	85.0	8,492.7	79.4	0.009	83.9	8,482.4	79.4	0.009	83.9

A5

HOUSEHOLDS WITH HEADS MOVED IN LAST 5 YEARS IN SAME MUNICIPALITY

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.27	1.82	0.49	0.24	1.68	0.27

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	571.5	28.5	0.050	577.3	27.1	0.047	95.2	569.4	26.7	0.047	93.8	568.0	26.5	0.047	93.0	569.0	26.5	0.047	93.0
9	370.4	23.8	0.064	372.5	23.1	0.062	96.9	376.5	22.3	0.059	93.3	378.4	22.2	0.059	93.2	377.5	22.2	0.059	93.0
11	339.8	22.2	0.065	339.9	22.0	0.065	99.1	333.7	21.3	0.064	95.9	335.3	21.3	0.063	95.9	334.7	21.3	0.064	95.9
2	310.7	21.3	0.069	310.6	21.4	0.069	100.3	312.8	21.1	0.068	99.2	312.5	21.1	0.068	99.2	312.6	21.1	0.068	99.2
12	94.4	12.5	0.133	93.5	12.5	0.133	99.5	94.5	12.2	0.129	97.4	94.3	12.2	0.129	97.4	94.3	12.2	0.129	97.4
ED	4,911.0	84.6	0.017	4,930.9	83.1	0.017	98.2	4,954.3	80.9	0.016	95.7	4,942.3	80.7	0.016	95.4	4,948.6	80.7	0.016	95.4

A6

HIGHEST GRADE OF HEAD IS 1 to 10

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
0.57	0.92	0.23	0.17	0.75	0.10

8	1,091.8	31.0	0.028	1,090.2	30.9	0.028	99.6	1,087.5	29.8	0.027	95.9	1,090.0	29.8	0.027	96.0	1,088.4	29.8	0.027	95.9
9	632.4	27.5	0.044	630.0	27.4	0.043	99.6	634.8	26.7	0.042	97.1	634.0	26.7	0.042	97.1	634.5	26.7	0.042	97.1
11	460.9	24.2	0.052	459.2	24.0	0.052	99.3	465.4	23.1	0.050	95.7	465.4	23.1	0.050	95.7	466.1	23.1	0.050	95.7
2	365.0	22.3	0.061	364.6	22.3	0.061	100.0	364.1	21.0	0.058	94.3	363.9	21.0	0.058	94.4	363.6	21.0	0.058	94.3
12	309.6	17.5	0.057	310.2	17.1	0.055	97.7	315.2	16.3	0.052	93.0	313.8	16.3	0.052	92.9	314.1	16.3	0.052	92.9
ED	9,386.6	93.9	0.010	9,355.4	93.4	0.010	99.5	9,356.8	89.7	0.010	95.5	9,362.0	89.7	0.010	95.5	9,357.6	89.7	0.010	95.4

A7

HEADS WITH BACHELOR DEGREE OR HIGHER

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
0.68	2.69	0.31	0.25	2.68	0.30

8	66.2	11.3	0.171	66.7	11.5	0.172	101.1	68.3	11.4	0.166	100.3	68.3	11.4	0.166	100.3	68.5	11.4	0.166	100.3
9	159.6	17.0	0.106	160.1	17.1	0.107	100.5	158.9	16.7	0.105	98.4	158.8	16.7	0.105	98.4	158.7	16.7	0.105	98.4
11	203.9	18.4	0.090	203.9	18.4	0.090	100.2	199.1	18.2	0.091	99.1	198.7	18.2	0.091	99.0	198.5	18.2	0.092	99.0
2	102.6	13.7	0.134	102.3	13.7	0.134	100.1	100.5	13.5	0.135	98.8	100.4	13.5	0.135	98.8	100.5	13.5	0.135	98.7
12	35.4	8.1	0.228	35.2	8.0	0.227	99.0	34.9	7.9	0.228	98.3	35.3	7.9	0.225	98.3	35.2	7.9	0.226	98.3
ED	1,304.3	48.6	0.037	1,306.0	48.6	0.037	100.0	1,316.8	47.7	0.036	98.2	1,317.2	47.7	0.036	98.1	1,318.4	47.7	0.036	98.1

B1

HOUSEHOLDS WITH 3 OR 4 PERSONS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
0.51	2.76	0.35	0.33	2.52	0.11

8	577.5	28.6	0.049	577.2	28.4	0.049	99.3	547.4	21.3	0.039	74.5	550.1	21.4	0.039	74.9	546.8	21.3	0.039	74.4
9	557.2	26.8	0.048	555.9	26.8	0.048	99.9	555.3	22.4	0.040	83.6	555.2	22.5	0.040	83.7	555.6	22.4	0.040	83.5
11	517.1	24.7	0.048	515.0	24.6	0.048	99.4	502.3	21.2	0.042	85.6	501.9	21.2	0.042	85.7	502.1	21.2	0.042	85.6
2	476.6	23.5	0.049	475.4	23.6	0.050	100.3	473.4	19.7	0.042	83.7	473.3	19.6	0.042	83.6	472.9	19.6	0.042	83.6
12	[illegible]	17.4	[illegible]	278.7	17.3	0.062	99.5	272.0	15.6	0.057	89.5	272.7	15.6	0.057	89.5	272.2	15.6	0.057	89.4
ED	6,983.4	93.4	0.013	6,977.1	92.8	0.013	99.3	6,910.7	79.0	0.011	84.6	6,921.5	79.2	0.011	84.8	6,908.8	79.0	0.011	84.6

CATEGORIES AT WA LEVEL

B2

AGE OF HEAD IS LESS THAN 25

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
3.76	6.13	1.65	1.01	5.12	0.68

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	255.7	21.1	0.083	260.4	20.3	0.078	96.3	260.9	17.0	0.065	80.3	257.4	17.0	0.066	80.7	259.9	16.9	0.065	80.1
9	102.4	13.9	0.136	104.7	13.3	0.127	95.4	105.6	12.4	0.118	89.3	107.0	12.4	0.116	89.3	106.2	12.3	0.116	88.9
11	44.3	9.2	0.207	45.3	8.6	0.190	94.0	42.1	8.3	0.197	90.5	43.2	8.2	0.190	89.4	42.8	8.2	0.191	89.4
2	96.5	13.3	0.138	96.5	12.8	0.133	96.1	97.9	11.8	0.120	88.0	97.0	11.8	0.121	88.2	97.4	11.7	0.120	87.9
12	32.4	7.8	0.240	32.7	7.7	0.235	99.1	30.6	7.3	0.239	94.2	31.1	7.3	0.235	94.1	31.0	7.3	0.235	94.1
ED	1,351.4	49.9	0.037	1,375.4	48.5	0.035	97.2	1,403.3	43.2	0.031	86.5	1,390.5	43.2	0.031	86.5	1,398.6	43.0	0.031	86.2

B3

AGE OF HEAD IS 25 TO 34

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.36	6.86	0.47	0.61	6.60	0.20

WA	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	324.8	23.3	0.072	327.9	23.2	0.071	99.3	306.7	15.9	0.052	68.1	305.2	16.1	0.053	68.9	306.9	15.9	0.052	68.0
9	334.3	23.0	0.069	334.7	22.2	0.066	96.3	326.8	12.4	0.038	54.0	328.4	12.9	0.039	56.0	326.5	12.4	0.038	53.9
11	322.1	21.8	0.068	320.7	21.6	0.067	99.1	278.9	8.3	0.030	38.1	280.4	8.4	0.030	38.8	278.4	8.3	0.030	37.8
2	307.7	21.2	0.069	305.8	20.3	0.067	95.8	306.2	11.3	0.037	53.0	305.6	11.5	0.038	54.3	306.5	11.2	0.037	52.9
12	188.7	16.1	0.085	187.9	16.0	0.085	99.5	174.0	6.9	0.040	43.1	174.6	7.0	0.040	43.4	173.6	6.9	0.040	43.1
ED	4,163.1	80.0	0.019	4,181.8	79.2	0.019	99.0	4,123.3	42.0	0.010	52.5	4,120.8	42.7	0.010	53.4	4,127.0	41.9	0.010	52.4

B4

HEADS WHO ARE WIDOWED, DIVORCED OR SEPARATED

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.40	4.75	0.64	0.57	4.45	0.12

WA	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	553.4	28.2	0.051	556.3	28.2	0.051	100.0	570.7	17.9	0.031	63.5	569.0	17.9	0.032	63.6	570.9	17.9	0.031	63.5
9	334.3	23.0	0.069	334.1	23.0	0.069	100.0	323.6	16.3	0.050	70.9	323.8	16.3	0.050	71.0	323.3	16.3	0.050	70.8
11	135.9	15.4	0.114	136.7	15.4	0.113	99.6	146.7	9.2	0.063	59.7	147.6	9.3	0.063	60.4	146.7	9.2	0.063	59.7
2	141.8	15.8	0.112	142.0	15.8	0.111	100.1	154.9	11.9	0.077	75.1	154.6	11.9	0.077	75.2	154.7	11.9	0.077	75.1
12	44.2	9.0	0.203	44.0	9.0	0.204	100.1	52.5	6.9	0.132	77.0	52.4	6.9	0.132	77.0	52.7	6.9	0.131	77.0
ED	3,177.3	71.7	0.023	3,189.3	71.2	0.022	99.3	3,278.4	49.0	0.015	68.4	3,266.5	49.3	0.015	68.7	3,276.9	49.0	0.015	68.3

C1

HOUSEHOLDS WITH 2 OR FEWER PEOPLE

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.36	4.45	0.66	0.60	4.05	0.07

WA	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est.	S.E.	C.V.	R.E.
8	1,037.7	31.2	0.030	1,042.2	29.6	0.028	94.8	1,091.1	10.7	0.010	34.2	1,084.8	11.4	0.011	36.5	1,090.7	10.6	0.010	34.1
9	614.4	27.4	0.045	617.2	26.4	0.043	96.4	613.6	10.5	0.017	38.3	615.0	10.9	0.018	39.8	613.7	10.5	0.017	38.2
11	384.1	23.0	0.060	388.3	22.0	0.057	95.4	392.2	8.4	0.021	36.3	393.6	8.8	0.022	38.0	392.2	8.3	0.021	36.2
2	365.0	22.3	0.061	366.8	22.0	0.060	98.9	375.7	9.0	0.024	40.4	374.9	9.1	0.024	41.0	375.8	9.0	0.024	40.4
12	165.1	15.4	0.093	163.5	15.3	0.094	99.4	169.8	8.2	0.048	52.9	169.4	8.2	0.048	53.1	170.1	8.2	0.048	52.9
ED	6,835.2	91.8	0.013	6,856.8	87.5	0.013	95.3	6,952.2	37.7	0.005	41.1	6,924.6	39.4	0.006	42.9	6,949.9	37.7	0.005	41.0

CATEGORIES AT WA LEVEL

C2

AGE OF HEAD IS 65 OR MORE

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.35	7.85	0.65	0.61	7.81	0.09

WA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est.	S.E.	C.V.	Pop.Est.	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.
8	472.2	26.8	0.057	469.8	26.5	0.056	98.7	486.5	6.2	0.013	23.2	487.3	6.4	0.013	23.8	486.8	6.2	0.013	23.2
9	259.0	20.9	0.081	257.5	20.5	0.080	98.5	266.1	4.7	0.018	22.3	265.1	4.7	0.018	22.7	265.7	4.6	0.018	22.3
11	109.3	14.0	0.128	109.5	14.1	0.128	100.4	110.3	7.2	0.065	51.5	110.1	7.2	0.065	51.4	110.0	7.2	0.065	51.2
2	111.6	14.3	0.128	113.5	14.4	0.127	101.4	103.2	2.4	0.023	16.6	103.4	2.4	0.023	16.8	103.2	2.4	0.023	16.6
12	32.4	7.8	0.240	32.8	7.8	0.238	100.6	42.6	4.1	0.096	53.0	42.1	4.1	0.098	53.1	42.4	4.1	0.097	53.0
ED	2,760.0	67.6	0.024	2,743.6	66.8	0.024	98.8	2,766.0	17.1	0.006	25.3	2,760.7	17.6	0.006	26.1	2,765.6	17.1	0.006	25.3

C3

OWNED DWELLINGS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
2.79	0.96	0.97	0.26	0.00	0.72

8	944.4	31.3	0.033	922.0	0.0	0.000	0.0	908.9	14.3	0.016	45.6	922.0	0.0	0.000	0.0	917.4	4.3	0.005	13.8
9	1,078.2	25.8	0.024	1,068.0	0.0	0.000	0.0	1,075.3	10.9	0.010	42.1	1,068.0	0.0	0.000	0.0	1,069.7	3.1	0.003	11.9
11	1,078.5	19.2	0.018	1,079.0	0.0	0.000	0.0	1,084.5	6.0	0.006	31.3	1,079.0	0.0	0.000	0.0	1,080.0	1.4	0.001	7.5
2	726.9	22.6	0.031	724.0	0.0	0.000	0.0	719.1	7.9	0.011	35.1	724.0	0.0	0.000	0.0	723.3	1.5	0.002	6.4
12	327.3	17.5	0.054	326.0	0.0	0.000	0.0	331.6	4.5	0.014	25.7	326.0	0.0	0.000	0.0	326.4	0.3	0.001	1.9
ED	12,624.0	89.3	0.007	12,477.0	0.0	0.000	0.0	12,419.8	34.7	0.003	38.9	12,477.0	0.0	0.000	0.0	12,459.2	9.3	0.001	10.4

C4

RENTED APARTMENTS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
5.46	3.27	3.20	0.85	0.15	2.41

8	649.7	29.5	0.045	661.0	0.0	0.000	0.0	680.8	11.7	0.017	39.7	661.0	0.0	0.000	0.0	665.2	3.4	0.005	11.6
9	286.1	21.7	0.076	299.0	0.0	0.000	0.0	294.1	8.7	0.030	40.1	299.0	0.0	0.000	0.0	297.7	2.4	0.008	11.2
11	82.7	12.3	0.149	90.0	0.0	0.000	0.0	87.7	5.6	0.064	45.5	90.0	0.0	0.000	0.0	89.2	1.4	0.016	11.8
2	84.5	12.6	0.149	95.0	0.0	0.000	0.0	95.1	3.3	0.035	26.4	95.0	0.0	0.000	0.0	95.2	0.4	0.005	3.5
12	20.6	6.3	0.303	20.7	6.2	0.299	98.9	20.0	6.0	0.300	96.0	20.4	6.0	0.294	96.0	20.4	6.0	0.294	96.0
RE	2,642.3	65.4	0.025	2,715.1	11.8	0.004	18.1	2,769.4	28.7	0.010	43.8	2,715.9	11.4	0.004	17.5	2,731.5	13.8	0.005	21.1

CATEGORIES AT EA LEVEL

A1

HOUSEHOLDS WITH EMPLOYED HEADS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.34	0.45	0.08	0.05	0.43	0.11

EA	Iteration 0			Iteration 1				Iteration 2				Iteration 3				Iteration 4			
	Pop.Est	S.E.	C.V.	Pop.Est	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.	Pop.Est	S.E.	C.V.	R.E.
2	428.3	23.1	0.054	423.4	21.9	0.052	94.9	424.0	21.3	0.050	92.1	424.1	21.2	0.050	91.9	424.3	21.2	0.050	91.9
1	362.0	22.2	0.061	360.0	21.4	0.059	96.3	362.5	20.9	0.058	93.8	362.3	20.8	0.058	93.7	362.1	20.8	0.058	93.7
3	129.7	15.2	0.117	135.3	14.7	0.108	96.8	135.3	14.2	0.105	93.2	135.7	14.1	0.104	92.6	135.7	14.1	0.104	92.6

A4

HOUSEHOLDS WITH HEADS NOT MOVED IN 5 YEARS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
0.24	2.14	0.48	0.20	1.99	0.28

2	117.6	14.6	0.124	118.0	14.6	0.123	99.8	121.0	14.2	0.118	97.6	121.4	14.2	0.117	97.5	121.3	14.2	0.117	97.5
1	171.9	17.1	0.100	172.2	16.3	0.095	95.3	168.9	15.7	0.093	91.6	169.8	15.6	0.092	91.1	169.4	15.6	0.092	91.1
3	15.1	5.5	0.363	15.2	5.5	0.362	100.4	14.8	5.5	0.370	100.0	14.9	5.5	0.368	100.0	14.9	5.5	0.368	100.0

A6

HIGHEST GRADE OF HEAD IS 1 TO 10

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.22	0.86	0.13	0.09	0.89	0.21

2	144.8	16.0	0.110	143.4	15.8	0.110	98.9	142.2	15.4	0.108	96.2	142.3	15.4	0.108	96.2	142.2	15.3	0.108	96.2
1	196.1	18.1	0.092	194.8	17.6	0.090	97.4	194.0	17.2	0.089	95.1	193.7	17.2	0.089	95.0	193.5	17.2	0.089	95.0
3	24.1	6.9	0.287	26.4	7.2	0.273	104.3	27.9	7.0	0.252	101.9	27.9	7.0	0.251	101.4	27.9	7.0	0.251	101.4

B1

HOUSEHOLDS WITH 3 OR 4 PERSONS

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
1.21	1.23	0.12	0.08	1.13	0.18

2	193.0	18.0	0.093	191.2	17.5	0.092	97.6	189.3	16.4	0.087	91.5	189.3	16.4	0.087	91.3	189.2	16.4	0.087	91.3
1	220.2	18.9	0.086	218.3	18.4	0.084	97.6	220.5	17.6	0.080	93.2	220.2	17.6	0.080	93.2	220.0	17.6	0.080	93.1
3	63.3	11.0	0.174	65.8	11.2	0.171	102.1	63.6	10.9	0.171	99.2	63.8	10.9	0.171	99.0	63.8	10.9	0.171	99.0

B3

AGE OF HEAD IS 25 TO 34

C(0-1)	C(1-2)	C(2-3)	C(3-4)	C(1-3)	C(2-4)
2.30	1.69	0.43	0.32	1.32	0.37

2	156.8	16.5	0.105	153.7	15.1	0.098	91.6	151.9	12.8	0.084	77.3	152.0	12.8	0.084	77.3	152.4	12.7	0.084	77.1
1	96.5	13.3	0.138	95.1	12.7	0.134	95.3	97.7	11.6	0.119	87.0	96.9	11.6	0.120	87.2	97.3	11.6	0.119	87.0
3	54.3	10.2	0.188	57.1	10.4	0.183	102.1	56.6	9.6	0.170	94.1	56.7	9.6	0.169	93.8	56.9	9.6	0.168	93.6

THE RESPONSE INCENTIVES EXPERIMENT IN THE CANADIAN LABOUR FORCE SURVEY

A.R. Gower
Household Surveys Development Division

This paper describes the methodology of the Response Incentives Experiment which was carried out in the Canadian Labour Force Survey in order to determine the effectiveness of a response incentive on improving respondent relations and interviewer performance. Included in the paper are various results relating to non-response rates and refusal rates as well as results of an evaluation questionnaire which was completed by all interviewers at the conclusion of the experiment.

1. INTRODUCTION

The Response Incentives Experiment was carried out in the Canadian Labour Force Survey during 1975 and 1976 in order to determine the effectiveness of a response incentive on improving respondent relations and interviewer performance.

The response incentive used in the experiment was the "Canada Handbook", an annual Statistics Canada publication which is an attractive presentation in textual and pictorial format of economic, social and cultural developments in Canada. The "Canada Handbook" was chosen because it was felt that this publication would improve respondent relations by showing respondents the importance of Statistics Canada's role as a compiler of statistical data and that it would provide interviewers with a valuable source of information on the various statistics produced by the agency for which they work.

This paper deals with the methodology and results of the Response Incentives Experiment. A comprehensive set of tables and graphs highlighting the major findings of the experiment is included. Section 2 describes the methodology of the experiment, including its design and the procedures

which were followed. Various results relating to non-response and refusal rates were obtained in order to determine the effect of the "Canada Handbook" on respondent relations and interviewer performance. These results are summarized in Section 3. Qualitative information on the effectiveness of response incentives in terms of interviewers' acceptance and attitudes was obtained from an evaluation questionnaire which was completed by all interviewers after the experiment was completed. Section 4 outlines the results of this evaluation questionnaire. Concluding remarks are given in Section 5.

2. METHODOLOGY OF THE EXPERIMENT

In order to measure the effectiveness of the "Canada Handbook" as a response incentive, interviewers across Canada were divided into two groups to form an experimental subsample and a control subsample. Interviewers in the experimental group distributed "Canada Handbooks" to all households which were visited for the first time, while interviewers in the control group offered no response incentive to respondents.

The two groups of interviewers were selected in such a way that the two groups together included all Labour Force Survey interviewers but did not overlap. The selection scheme was carried out independently for SRU interviewers and NSRU interviewers. SRU interviewers interview primarily in SRU's (self-representing units, i.e. cities with a population of over 12,000), while NSRU interviewers interview primarily in NSRU's (non self-representing units, i.e. areas other than SRU's).

The first step in the selection procedure was to stratify the interviewers according to the eight regional offices (St. John's, Halifax, Montreal, Ottawa, Toronto, Winnipeg, Edmonton and Vancouver). Within each regional office, interviewers were assigned to the experimental and control groups in the following way: (a) If an SRU had four or more interviewers, then these interviewers were listed according to the magnitude of their average refusal rates over a three month period prior to the implementation

of the experiment (i.e. the interviewer with the lowest refusal rate was listed first, while the interviewer with the highest refusal rate was listed last). One of the first two interviewers on the list was then chosen at random, and this interviewer together with every second interviewer who followed was assigned to the experimental group. All remaining interviewers were assigned to the control group. (b) All SRU's having less than four interviewers were grouped together and interviewers in these areas were systematically allocated to one of the two groups using the method described in (a) above. (c) Interviewers in NSRU's were systematically allocated to the experimental or control group according to the procedure described in (a) above.

It should be pointed out that this method of subsample selection allowed the experiment to be easily implemented in the field, and it ensured that refusal rates in each subsample were approximately equal at the outset of the experiment.

Regional Offices were required to report all new hirings and changes in interviewer assignments to Head Office in order to facilitate the analysis of non-response rates. If an interviewer was replaced by another interviewer, then the procedure was to assign the new interviewer to whichever group the original interviewer belonged. In cases where interviewers were hired to enumerate areas with a sample increase (i.e. not a replacement), these interviewers were systematically allocated to the two groups.

The Labour Force Survey sample consists of six rotation groups, each of approximately equal size. Every selected dwelling belongs to one of these rotation groups and remains in the survey for six consecutive months. In any one month approximately one-sixth of the sample rotates out and is replaced by dwellings rotating into the sample for the first time (for example, a dwelling which rotates into the survey in January is enumerated each month from January to June and is replaced by another dwelling in July). Interviewers in the experimental group distributed

one copy of the "Canada Handbook" to each household entering the survey for the first time. A copy was given in any of the following circumstances:

(a) during the first interview, when a dwelling rotated into the survey for the first time.

(b) at the time of the first interview with a household which was a non-interview in the first month or in all previous months of its six month tenure in the survey, and

(c) when there was a complete change in household membership during the six months while a dwelling was in the survey.

Interviewers were instructed to give the "Canada Handbook" to the respondent at the end of the interview, unless they felt it would be more effective to offer it at the beginning of the interview (for example, if they detected some resistance on the part of the respondent). Interviewers explained to the respondents that the "Canada Handbook" was being given to them in appreciation for their co-operation during the survey. When confronted with a refusal on a first visit interviewers still offered a copy of the "Canada Handbook" to the respondent, with a brief explanation that these books were being distributed to every selected household.

Interviewers in the SRU experimental group began to distribute "Canada Handbooks" during the August 1975 survey to households visited for the first time. Most of these households, it should be pointed out, were households rotating into the Labour Force Survey sample that month. SRU interviewers continued to distribute "Canada Handbooks" until the January 1976 survey. In this way the response incentive was eventually distributed to all households in their assignments which were contacted for the first time during the six month period from August to January. The control subsample, of course, received no response incentive over this same period of time other than the usual introductory letter and presentation of the interviewer's identification card.

Because of problems (such as assignments being re-distributed and new interviewers being hired) caused by an increase in sample size in the NSRU portion of the Labour Force Survey, it was decided to implement the Response Incentives Experiment at a different time in NSRU's than in SRU's. For this reason, the experiment did not begin in NSRU's until the October 1975 survey, and it ran until the March 1976 survey. Every household which was enumerated for the first time over this six month period by an interviewer in the NSRU experimental group received a copy of the "Canada Handbook" and, like the SRU control group, no response incentive was distributed by interviewers in the NSRU control group.

3. NON-RESPONSE AND REFUSAL RATES

In the Labour Force Survey, non-response occurs due to operational difficulties, time and cost restraints, the lack of co-operation from respondents, the inability or unwillingness of interviewers to track down missing respondents, or for some other reason. The non-response rate measures the severity of this non-response problem, and it is calculated as the percentage of non-respondent households out of all sampled households.

The non-response rates (including all components of non-response) and refusal rates were averaged over the six rotation groups on the basis of the number of months each group of households was in the survey. These rates were calculated for the control and experimental subsamples in the following way. Let R_j denote the average non-response rate for households during the jth month they were in the survey. Then

$$R_j = \frac{\sum_{i=1}^{6} r_{ij}}{\sum_{i=1}^{6} n_{ij}} \times 100 \qquad (j = 1,2,3,4,5,6)$$

where r_{ij} and n_{ij} denote the number of non-responses and the number of sampled households respectively in rotation group i during the jth month of the rotation pattern. The average refusal rate was calculated in a similar fashion.

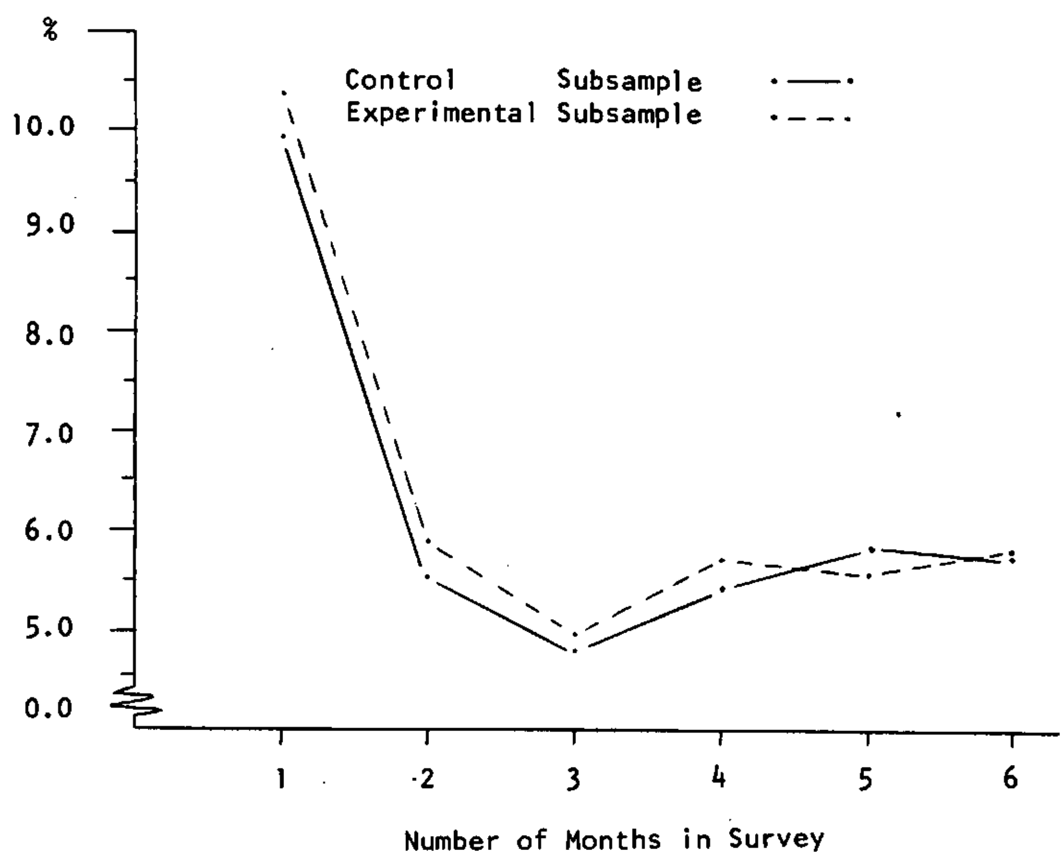
To test for significant differences in the rates (non-response or refusal) between the two subsamples the statistic

$$t = \frac{p_c - p_e}{\sqrt{\frac{p_c(1 - p_c)}{n_c} + \frac{p_e(1 - p_e)}{n_e}}}$$

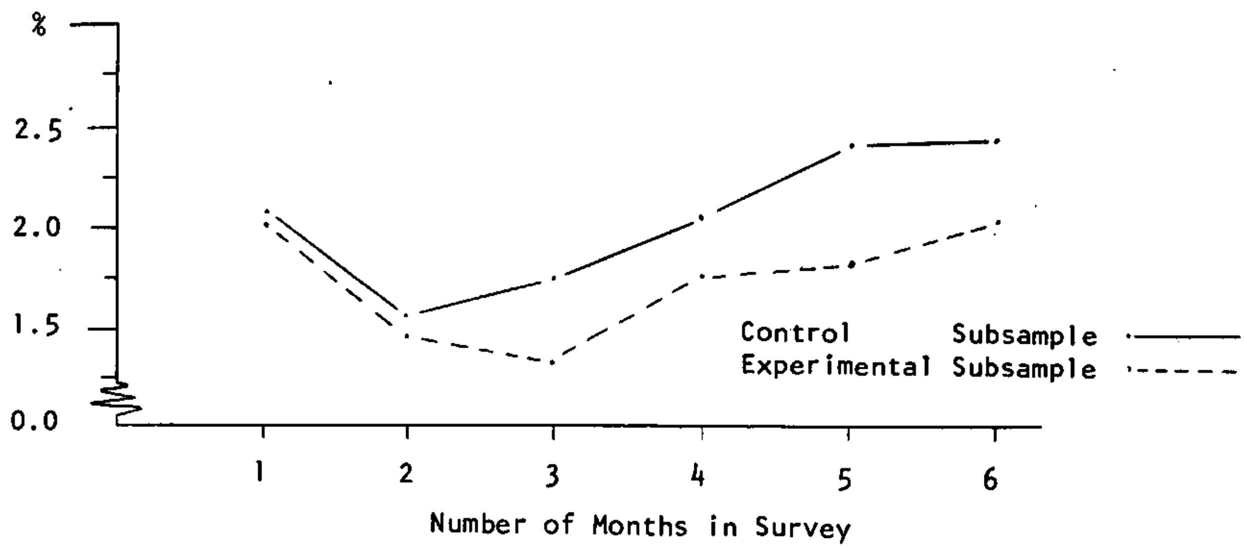
was calculated, where p_c and p_e denote the rates in the control and experimental subsamples respectively and n_c and n_e denote the number of households in the control and experimental subsamples respectively. If $|t|$ exceeded 1.96, then the difference in the rates between the two subsamples was considered to be significant at the 5 percent level of significance.

Graphs 3.1 and 3.2 show the non-response and refusal rates respectively for SRU's, averaged over the six rotation groups.

Graph 3.1: Non-Response Rates in the SRU Control and Experimental Subsamples Averaged Over Six Rotation Groups



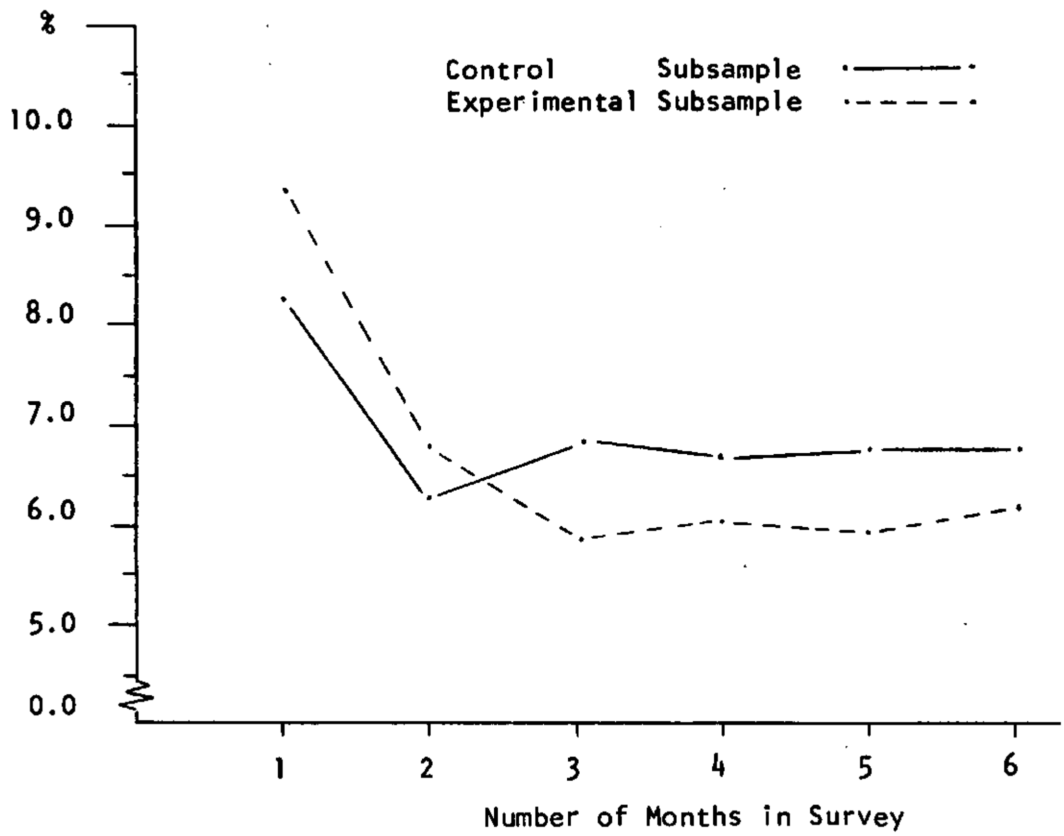
Graph 3.2: Refusal Rates in the SRU Control and Experimental Subsamples Averaged Over Six Rotation Groups



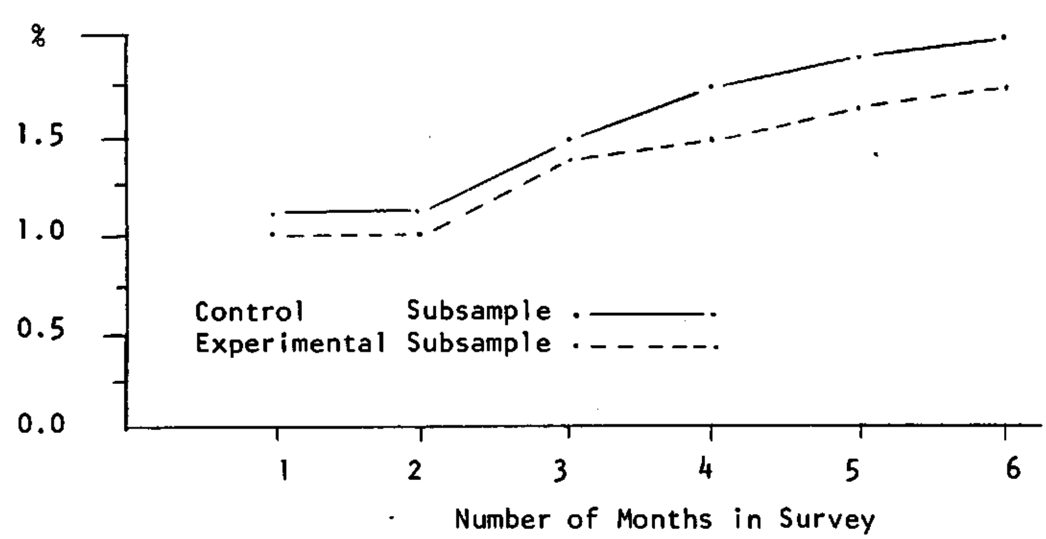
Graph 3.1 shows that the average non-response rates in the SRU control and experimental subsamples were approximately equal during every month of the six-month rotation pattern. No significant differences between the two subsamples were noted in the average non-response rates. In the case of the average refusal rates, Graph 3.2 shows that the refusal rate in the SRU experimental subsample was consistently lower than the corresponding rate in the SRU control subsample throughout the six months, with the differences being significant in the third and fifth months only.

Graphs 3.3 and 3.4 show the average non-response and refusal rates respectively for NSRU's.

Graph 3.3: Non-Response Rates in the NSRU Control and Experimental Subsamples Averaged Over Six Rotation Groups



Graph 3.4: Refusal Rates in the NSRU Control and Experimental Subsamples Averaged Over Six Rotation Groups



In NSRU's Graph 3.3 shows that the average overall non-response rate for the six rotation groups was higher in the experimental subsample than in the control subsample for the first two months but lower during the remaining four months. These differences were significant in the first, third and fifth months.

Like the behaviour of the SRU refusal rates, Graph 3.4 shows that the refusal rate in the NSRU experimental group was consistently lower than the refusal rate in the NSRU control group. No significant differences, however, were observed in any of the six months.

The behaviour of refusal rates in the control and experimental subsamples in both SRU's and NSRU's indicates that the distribution of the "Canada Handbook" had very little effect on converting a refusal to a response at the time of the interviewer's first contact with a household, but respondents who received the "Canada Handbook" were less likely to refuse at some later time than respondents who received no response incentive. The increasing trend which was observed in the behaviour of refusal rates in the two subsamples during the six month tenure of households in the survey was due to the fact that refusal rates tend to be cumulative. A refusal one month usually remains a refusal the next month, so that an increase in the refusal rate during one month can be expected to result in higher refusal rates during subsequent months.

During the experiment, it was found that other types of non-response, such as the "no one at home" component, were generally not any lower in the experimental subsample than in the control subsample. This suggests that having the "Canada Handbook" available to give to respondents did not motivate interviewers to put more effort into tracking down non-respondents.

4. EVALUATION QUESTIONNAIRE

Interviewers in both the experimental and control groups were asked to complete an evaluation questionnaire in order that qualitative information could be obtained on the effectiveness of response incentives in terms of interviewers' acceptance and attitudes. Interviewers rated their own attitudes as well as their perception of respondent attitudes to a list of miscellaneous materials which covered a wide range of response incentives and represented varying degrees of association with Statistics Canada and the Labour Force Survey. Included on this list were the following items:

(1) the Labour Force Survey brochure - a pamphlet outlining the purpose of the Labour Force Survey and asking respondents for their co-operation,

(2) an interviewer identification card, with a photograph of the interviewer attached, that shows the interviewer to be an official representative of Statistics Canada (to be shown to respondents),

(3) a letter from Statistics Canada to respondents prior to the first interview that explains the importance of the Labour Force Survey and asks for their co-operation,

(4) a Statistics Canada publication that describes the use of Labour Force Survey data and other data collected by Statistics Canada in a colour-illustrated paperback format ("Canada Handbook"),

(5) a reference telephone number to be given to respondents who have questions about the Labour Force Survey which cannot be answered by the interviewer,

(6) a metric converter that gives conversions of temperature and other measures to the metric scale (e.g. Fahrenheit to Centigrade),

(7) a wallet size calendar, and

(8) nothing necessary.

Interviewers were asked to show the extent of their agreement or disagreement on a scale of 1 to 5 (strongly disagree to strongly agree) to the following two statements which describe the effect of the response incentives listed above on the respondent (statement A) and on the interviewer (statement B).

Statement A: "Respondents would be more willing to participate in the Labour Force Survey if they were given such a material".

Statement B: "You would find it easier to ask respondents to participate if they were given such a material".

In addition to completing these questions, interviewers in the experimental group were also asked to respond to the following questions:

(1) How did you find respondents generally reacted to the "Canada Handbook"? (a) very positive, (b) positive, (c) neutral, (d) negative or (e) very negative.

(2) If the "Canada Handbook" or some other publication was distributed to respondents, to which households do you think it should be sent? (a) all households, (b) hostile households, or (c) other (specify).

(3) If the "Canada Handbook" or some other publication was distributed to respondents, when do you think it should be distributed? (a) before the first interview, sent by the regional office, (b) first interview, (c) anytime, (d) last interview, or (e) after the last interview, sent by the regional officc.

(4) If the "Canada Handbook" or some other publication was distributed during the interview, it should be ... (a) at the beginning of the interview, (b) as decided by the interviewer, or (c) upon completion of the interview.

(5) Was the distribution of the "Canada Handbook" difficult to manage while doing your assignment? (a) yes, or (b) no.

A total of 552 interviewers across Canada completed the response incentives evaluation questionnaire. These 552 interviewers can be broken down as follows:

126	interviewers	in the	SRU	control group
150	"	" "	NSRU	" "
128	"	" "	SRU	experimental group
148	"	" "	NSRU	" "

Unfortunately, no responses were received from 100 interviewers (50 in the control group and 50 in the experimental group).

The results of the evaluation questionnaire are summarized on Table 4.1 for the control and experimental groups for both SRU's and NSRU's combined. The median rating has been used to summarize the responses to each statement or question. When asked to show the extent of their agreement or disagreement with statements A and B describing the effect of the various response incentives on respondents as well as themselves, interviewers indicated

- strong agreement with the letter
- agreement with the interviewer identification card, Labour Force Survey brochure, "Canada Handbook" and reference telephone number
- uncertainty about the wallet-size calendar and metric converter
- disagreement that nothing was necessary.

The frequencies of the ratings for each statement on the questionnaire were also tabulated for interviewers according to the following classifications:

(1) type of area (SRU or NSRU)
(2) interviewing experience, and
(3) size of assignment (i.e. number of dwellings assigned).

The reaction of SRU interviewers to the various response incentives was very similar to the reaction of NSRU interviewers. Interviewers in SRU's, however, tended to show stronger agreement with the Labour Force Survey brochure and the reference telephone number than did interviewers in NSRU's. In both SRU's and NSRU's most interviewers indicated very strong agreement with the introductory letter and lesser agreement, in varying degrees, with the other response incentives. Differences in the responses by SRU and NSRU interviewers may, in part, have been due to the effect of the new interviewers who were hired to enumerate areas in NSRU's where there was a sample size increase. As pointed out in Section 2, however, precautions were taken to ensure that new interviewers were included in both the control and experimental groups by systematically allocating them to the two groups.

Table 4.1: RESULTS OF LFS RESPONSE INCENTIVES EXPERIMENT EVALUATION QUESTIONNAIRE

(Completed by all SRU and NSRU interviewers across Canada)

Statement A: Respondents would be more willing to participate in the LFS if they were given . . .

Statement B: You (the interviewer) would find it easier to ask respondents to participate in the LFS if they were given . . .

(1) the LFS brochure

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	1	4	18	36	41	276	Agree
	Experimental	1	1	22	34	42	276	Agree
B	Control	1	2	18	34	45	276	Agree
	Experimental	1	1	18	33	47	276	Agree

(2) an interviewer identification card

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	1	8	14	36	41	276	Agree
	Experimental	1	9	14	37	39	276	Agree
B	Control	1	9	14	34	42	276	Agree
	Experimental	2	10	15	36	37	276	Agree

(3) an introductory letter

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	2	1	5	24	68	276	Strongly Agree
	Experimental	1	2	10	21	66	276	Strongly Agree
B	Control	2	2	4	21	71	276	Strongly Agree
	Experimental	1	1	9	24	65	276	Strongly Agree

(4) a Statistics Canada publication (e.g. "Canada Handbook")

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	1	8	28	30	33	276	Agree
	Experimental	1	6	15	38	40	276	Agree
B	Control	2	12	22	34	30	276	Agree
	Experimental	2	7	13	38	40	276	Agree

Table 4.1 (continued)

Statement A: Respondents would be more willing to participate in the LFS if they were given . . .

Statement B: You (the interviewer) would find it easier to ask respondents to participate in the LFS if they were given . . .

(5) a reference telephone number

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	4	17	27	32	20	276	Agree
	Experimental	6	14	28	34	18	276	Agree
B	Control	4	15	29	32	20	276	Agree
	Experimental	5	19	24	32	20	276	Agree

(6) a metric converter

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	12	32	26	19	11	276	Uncertain
	Experimental	15	21	31	21	12	276	Uncertain
B	Control	14	33	26	18	9	276	Uncertain
	Experimental	15	24	28	24	9	276	Uncertain

(7) a wallet-size calendar

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	14	30	27	22	7	276	Uncertain
	Experimental	15	28	33	18	6	276	Uncertain
B	Control	15	33	26	19	7	276	Uncertain
	Experimental	14	28	34	18	6	276	Uncertain

(8) nothing necessary

Statement	Group	Percent Responses: Strongly Disagree	Disagree	Uncertain	Agree	Strongly Agree	No. of Responses	Median Response
A	Control	41	27	18	11	3	276	Disagree
	Experimental	39	28	20	10	3	276	Disagree
B	Control	40	25	21	10	4	276	Disagree
	Experimental	40	25	22	10	3	276	Disagree

When classified according to interviewing experience, interviewers with at least one year's experience tended to show stronger agreement with the interviewer identification card than did interviewers with less than one year's experience. In fact, 43 percent of the interviewers with at least one year's experience indicated "strong agreement" with the identification card, while only 28 percent of the remaining interviewers indicated "strong agreement" with this response incentive. The extent of interviewers' agreement or disagreement with the other response incentives listed on the questionnaire did not vary according to the number of months of interviewing experience. Similarly, questionnaire responses did not vary according to the size of interviewers' assignments, except that interviewers with larger assignments (60 dwellings or more) tended to show stronger agreement with the Labour Force Survey brochure, interviewer identification card and letter but slightly less agreement with the "Canada Handbook" than interviewers with smaller assignments (less than 60 dwellings).

Responses by interviewers in the experimental group to the five questions relating to the distribution of the "Canada Handbook" are summarized in Table 4.2. The percent responses and the median response to each question are given in the table.

Most interviewers (87 percent) rated respondents' reaction to the "Canada Handbook" as "positive" or "very positive", while only a few interviewers (under 2 percent) felt that respondents' reaction was "negative" or "very negative". The median response of all interviewers was that the reaction of respondents was "very positive". Interviewers in SRU's differed slightly in their assessment of respondents' reaction than interviewers in NSRU's. SRU interviewers tended to rate respondents' reaction more positively than did NSRU interviewers and, whereas the median response in SRU's was "very positive", the median response in NSRU's was "positive". Similarly, interviewers with at least one year's experience rated the reaction of respondents as "very positive", while interviewers with less than one year's experience rated their reaction as "positive".

Table 4.2: SUMMARY OF RESULTS OF RESPONSE INCENTIVES EVALUATION QUESTIONNAIRE

Questions completed by all interviewers across Canada in the experimental group:

Question 1: How did you find respondents generally reacted to the "Canada Handbook"?

Percent Responses					No. of Responses	Median Response
Very Negative	Negative	Neutral	Positive	Very Positive		
1	1	11	36	51	270	Very Positive

Question 2: If the "Canada Handbook" or some other publication was distributed to respondents, to which households do you think it should be distributed?

Percent Responses			No. of Responses	Median Response
All Households	Hostile Households	Other		
86	3	11	269	All Households

Question 3: If the "Canada Handbook" or some other publication was distributed to respondents, when do you think it should be distributed? (A) Before first interview, sent by regional office, (B) first interview, (C) anytime, (D) last interview, (E) after last interview, sent by regional office.

Percent Responses					No. of Responses	Median Response
A	B	C	D	E		
14	76	3	6	1	270	1st Interview

Question 4: If the "Canada Handbook" or some other publication was distributed during the interview, it should be . . .

Percent Responses			No. of Responses	Median Response
At the beginning of the interview	As decided by the interviewer	Upon completion of the interview		
24	54	22	271	As decided by the interviewer

Question 5: Was the distribution of the "Canada Handbook" difficult to manage while doing your assignment?

Percent Responses		No. of Responses	Median Response
Yes	No		
12	88	268	No

Nearly every interviewer replied that the "Canada Handbook" or some other response incentive should be given to all households. Most interviewers felt that this type of response incentive should be given to the respondent at the time of the first interview. Some interviewers indicated that they would prefer to have it sent to the respondent by the regional office prior to the first interview, but very few interviewers thought that the last interview was the appropriate time.

When queried on the best time during the interview to give the "Canada Handbook" to the respondent, slightly more than half the interviewers felt that it should be left to the discretion of the interviewer, while the remaining interviewers were almost evenly divided between the responses "at the beginning of the interview" and "upon completion of the interview". Interviewers with more experience tended to choose "upon completion of the interview" more frequently than interviewers with less experience.

Almost 90 percent of the interviewers who distributed the "Canada Handbook" responded that they did not find it difficult to manage while doing their assignments. The remaining 10 percent of the interviewers indicated that they found the book difficult to manage. Some of these interviewers wrote comments, saying that the books were heavy and difficult to carry in and out of their cars and home. Interviewers who had large assignments apparently found it more difficult than interviewers with smaller assignments to distribute copies of the "Canada Handbook" while doing their assignments. About 21 percent of the interviewers with assignment sizes greater than 60 dwellings responded that they found it difficult, while only 6 percent of the interviewers with assignment sizes less than 60 dwellings found the distribution of "Canada Handbooks" difficult to manage.

Interviewers wrote many comments in addition to their responses to the evaluation questionnaire. Most interviewers responded that the distribution of the "Canada Handbook" was very worthwhile and that it was especially well received by professionals, students and respondents with

school-age children. Some interviewers, however, encountered respondents who considered that the "Canada Handbook" was an expensive publication to be given out so generously and that it was a further example of the waste of taxpayers' money. A few interviewers found that presenting the "Canada Handbook" to hostile respondents often made them more hostile. Many interviewers indicated that a less expensive response incentive is probably all that is necessary, and they suggested that a small publication relating only to the survey data and its uses should be given to respondents. Most interviewers feel that it is necessary to educate respondents as well as the general public on the purposes and uses of the survey. For this reason, interviewers feel that it is extremely important to send an introductory letter to respondents in order to explain the purpose of the survey and to ask for their co-operation. Some interviewers also suggested that a "thank you" letter should be sent to the respondents after the survey thanking them for their participation. Generally speaking, it can be concluded that the majority of interviewers feel that response incentives are necessary and that they are helpful in establishing a good rapport with respondents.

5. CONCLUDING REMARKS

The Response Incentives Experiment provided useful information on the effect of a response incentive such as the "Canada Handbook" on respondent relations and interviewer performance.

Households which received the "Canada Handbook" tended to have a lower refusal rate than households which did not receive it, although the observed differences in the refusal rates between the control and experimental groups were generally not significant. The behaviour of refusal rates in the two groups indicated that the distribution of the "Canada Handbook" had very little effect on converting a refusal to a response at the time of the interviewer's first contact with a household, but respondents who received the "Canada Handbook" were less likely to refuse at some later time than respondents who received no response incentive. Considering that the refusal rate in the Labour Force Survey

under normal survey conditions is very low (approximately 1.5 percent), it is not surprising that only a minimal reduction in the refusal rate was realized through the distribution of the "Canada Handbook". Other types of non-response, such as "no one at home", were generally higher in the experimental subsample than in the control subsample. This suggests that the "Canada Handbook" did not motivate interviewers to put more effort into tracking down non-respondents.

The results of the evaluation questionnaire showed that both respondents and interviewers reacted very favourably to the "Canada Handbook". Although the majority of interviewers feel that response incentives are useful in establishing a good rapport with respondents, most interviewers believe that materials such as an introductory letter and an identification card are actually more effective than gifts such as the "Canada Handbook".

The Response Incentives Experiment has shown that there is a real need to provide the respondent with more information on the purposes of the survey and the uses of the data. It is very important, therefore, that interviewers should be equipped to provide this information since they have the main responsibility in gaining the co-operation of the respondent. Interviewers can be equipped with this knowledge through training which emphasizes the purposes and importance of the survey and by having support material available such as an introductory letter, an explanatory brochure or other response incentives which illustrate these points.

6. ACKNOWLEDGMENT

Acknowledgment is due to the HSDS-Field Research Steering Committee consisting of Mr. P.F. Timmons, Mr. R.C. Muirhead, Mr. H. O'Haver and Mr. E.T. McLeod under whose direction the Response Incentives Experiment was carried out. The contribution of Mr. J. Pye and other staff members of Field Division in undertaking the experiment is also acknowledged. The author wishes to express his appreciation to the referee and the editors for their very helpful comments.

RESUME

Cet article présente une description de la méthodologie de l'expérience sur l'incitation à la réponse qui a été effectuée à l'occasion de l'enquête sur la population active du Canada afin de déterminer comment une incitation à la réponse pourrait améliorer les rapports entre le répondant et l'enquêteur et le rendement de ce dernier. On y retrouvera aussi divers résultats concernant les taux de non-réponse et de refus, de même que les réponses au questionnaire d'évaluation qu'ont rempli tous les enquêteurs à la fin de l'expérience.

ESTIMATION OF VARIANCE IN MULTI-STAGE SAMPLING

K.P. Srinath and M.A. Hidiroglou
Business Survey Methods Division

In multi-stage sampling when selection is without replacement at the first stage, estimation of the variance of the estimate of the population total is often done assuming sampling with replacement. This estimate is biased and the degree of bias is not negligible. In this paper, a procedure which gives unbiased estimates of the variance making use of only estimated primary sampling unit totals is suggested for the case when sampling at the second and subsequent stages is simple random without replacement. This procedure is based on sub-samples drawn from the selected second and subsequent stage units.

1. INTRODUCTION

In multi-stage sampling when selection is without replacement at the first stage, unbiased estimation of the variance of the estimate of the population total involves the computation of within primary sampling unit (PSU) variances. The computation of these variances can be tedious and costly. Often, in practice, a less rigorous but simpler biased estimate of the variance is obtained by assuming sampling with replacement at the first stage. This bias could be quite large under certain situations and therefore, it may be of interest to look for a procedure which is simple and which gives unbiased estimates.

In this note such a procedure is suggested for the case when sampling at the second and subsequent stages is simple random without replacement. In this procedure a sub-sample of units is selected from the sample of second and subsequent stage units according to a suggested sub-sampling rule. A new set of PSU totals is obtained based on the selected sub-samples. An unbiased estimator of the variance of the estimator of the population total is then obtained by simply considering the variance between the estimated PSU totals based on the selected sub-samples. This

procedure may be useful when a quick estimate of the variance is required. In this paper, the bias of the variance estimate assuming sampling with replacement is also examined. Some comparisons of the suggested estimator with the usual unbiased estimator and the biased estimator are made.

2. VARIANCE ESTIMATOR

A sample of n PSU's is selected from N PSU's without replacement and with unequal probabilities. Let π_i be the probability that the ith PSU U_i is in the sample. Let π_{ij} be the probability that both U_i and U_j are in the sample. Let T_i be an unbiased estimator of ith PSU total Y_i based on sub-sampling at the second and subsequent stages. Sampling in these stages is assumed to be simple random without replacement.

An unbiased estimator of the population total is given by

$$\hat{Y} = \sum_{i=1}^{n} \frac{T_i}{\pi_i} \tag{2.1}$$

and the variance of $\hat{Y}$ is given by

$$V(\hat{Y}) = \sum_{i<j}^{N}\sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i} \tag{2.2}$$

where σ_i^2 is the conditional variance of T_i whatever be the number of stages. The usual unbiased variance estimator of (2.2) is

$$v_1(\hat{Y}) = \sum_{i<j}^{n}\sum \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}}\right) \left(\frac{T_i}{\pi_i} - \frac{T_j}{\pi_j}\right)^2 + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} \tag{2.3}$$

where $\hat{\sigma}_i^2$ is an unbiased estimator of σ_i^2.

The biased estimator of $V(\hat{Y})$ assuming sampling with replacement at the first stage is

$$v_2(\hat{Y}) = \frac{1}{n^2(n-1)} \sum_{i<j}^{n}\sum \left(\frac{T_i}{P_i} - \frac{T_j}{P_j}\right)^2 \quad , \tag{2.4}$$

where P_i and P_j are the probabilities of selection of ith and jth PSU's in each of the draws. The bias in $v_2(\hat{Y})$ can easily be shown to be equal to

$$B[v_2(\hat{Y})] = \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i} \left(\frac{\pi_i^2}{n^2 P_i^2} - 1\right) + \sum_{i<j}^{N}\sum \left[\frac{\pi_{ij}}{n^2(n-1)} \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2\right.$$

$$\left. - (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2\right] . \tag{2.5}$$

For the special case $\pi_i = nP_i$ we have

$$B[v_2(\hat{Y})] = \sum_{i<j}^{N}\sum \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 \left(\frac{n\pi_{ij} - (n-1)\,\pi_i \pi_j}{(n-1)}\right) .$$

The above expression can be rewritten as

$$B[v_2(\hat{Y})] = \frac{n}{n-1} \left[\frac{1}{n} \sum_{i<j}^{N}\sum P_i P_j \left(\frac{Y_i}{P_i} - \frac{Y_j}{P_j}\right)^2 - \sum_{i<j}^{N}\sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2\right] . \tag{2.6}$$

That is, the bias in $v_2(\hat{Y})$ is $\frac{n}{n-1}$ times the reduction in variance (if any) obtained by sampling without replacement instead of with replacement at the first stage, as shown by Durbin [3]. This bias may not be negligible especially when efficient procedures of without replacement sampling are used. An example showing the amount of this bias for certain characteristics has been given by Des Raj [1].

2.1 Sub-sampling Rule for Two-Stage Sampling

In case of two-stage sampling when selection is simple random without replacement at the second stage, we have $T_i = M_i \bar{y}_i$ where $\bar{y}_i$ is the mean based on m_i units drawn from M_i units belonging to the ith selected PSU.

$$\sigma_i^2 = M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \quad \text{and} \quad \hat{\sigma}_i^2 = M_i^2 \left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 \tag{2.7}$$

where S_i^2 is the mean square between the M_i secondary sampling units (ssu's) in the ith selected PSU and s_i^2 the mean square between the selected m_i ssu's in the ith selected PSU. In order to estimate (2.2), using our procedure, a simple random sub-sample of m_i' units is drawn without replacement from the m_i units in the ith selected PSU. The rule for determining m_i' in order to obtain an unbiased estimator of $V(\hat{Y})$ is given in (2.11).

An unbiased estimator of the variance of $\hat{Y}$ in (2.2) is given by

$$v_3(\hat{Y}) = \sum_{i<j}^{n}\sum \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{T_i'}{\pi_i} - \frac{T_j'}{\pi_j}\right)^2 \tag{2.8}$$

where $T_i' = M_i \bar{y}_i'$ and $\bar{y}_i'$ is the sample mean based on m_i' units. The proof of the result is as follows. Consider

$$E\, v_3(\hat{Y}) = E_1 E_2 \sum_{i<j}^{n}\sum \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} E_3 \left(\frac{T_i'}{\pi_i} - \frac{T_j'}{\pi_j}\right)^2 \tag{2.9}$$

where E_3 refers to the conditional expectation over all selections of $m_i', m_2', \ldots m_n'$ from $m_1, m_2, \ldots m_n$ which are kept fixed, E_2 is the conditional expectation over all selections of $m_1, m_2, \ldots m_n$ ssu's from the PSU's which are kept fixed and E_1 denotes the expectation over all possible samples of n PSU's from N. Now we have

$$E_3 \left(\frac{T_i'}{\pi_i} - \frac{T_j'}{\pi_j}\right)^2 = \left(\frac{T_i}{\pi_i} - \frac{T_j}{\pi_j}\right)^2 + \left(\frac{\sigma_i'^2}{\pi_i^2} + \frac{\sigma_j'^2}{\pi_j^2}\right)$$

where $\sigma_i'^2 = M_i^2 \left(\frac{1}{m_i'} - \frac{1}{m_i}\right) s_i^2$.

Next,

$$E_2\left[\left(\frac{T_i}{\pi_i} - \frac{T_j}{\pi_j}\right)^2 + \left(\frac{\sigma_i'^2}{\pi_i} + \frac{\sigma_j'^2}{\pi_j}\right)\right] = \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2$$

$$+ \frac{M_i^2}{\pi_i^2} \left(\frac{1}{m_i'} - \frac{1}{M_i}\right) S_i^2 + \frac{M_j^2}{\pi_j^2} \left(\frac{1}{m_j'} - \frac{1}{M_j}\right) S_j^2 .$$

Therefore,

$$E[v_3(\hat{Y})] = \sum_{i<j}^{N}\sum (\pi_i \pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N} \frac{M_i^2}{\pi_i} \left(\frac{1}{m_i'} - \frac{1}{M_i}\right) S_i^2 (1-\pi_i). \quad (2.10)$$

It is easily seen from (2.10) and (2.2) that by making

$$m_i' = \frac{m_i (1-\pi_i)}{1 - \pi_i \frac{m_i}{M_i}} \quad (2.11)$$

we get

$$E[v_3(\hat{Y})] = V(\hat{Y}) .$$

It is interesting to note that while it is not possible to obtain an unbiased estimator of the total variance using only PSU totals based on all the sampled ssu's, it is possible to obtain it using PSU totals based on a sub-sample of ssu's.

2.2 Sub-sampling Rules for Multi-Stage Sampling

Sub-sampling rules for multi-stage sampling are a straight forward extension of the rule for two-stage sampling. If the design has u stages, the number of νth stage units to be sub-sampled from the sample of νth stage units in the ith selected PSU for the purpose of variance estimation is given by

$$q'_{ijk.....\nu} = \frac{q_{ijk.....\nu}\left(1 - \pi_i \prod_{\ell=2}^{\nu-1} f_{i\ell}\right)}{1 - \pi_i \prod_{\ell=2}^{\nu} f_{i\ell}} \qquad \nu = 2, \ldots u$$

where $q_{ijk...\nu}$ is the number of νth stage units in the sample and $f_{i\ell}$ ($\ell = 2,3, \ldots\nu$) is the sampling fraction at the ℓth stage in the ith PSU.

3. EMPIRICAL STUDY

To investigate the efficiency of the proposed estimator relative to the biased estimator and the usual unbiased estimator, we have chosen two artificial populations of households. These were taken from Som [6]. Tables 1 and 2 give the values of M_i, Y_i, P_i, m_i and S^2_{wi} for each of the PSU's in each of the two populations. Individual values of y_{ij} are not shown.

We carried out a limited sampling experiment in view of the large number of possible samples of ssu's that could be drawn from the PSU's. Two samples of ssu's, each of size m_i were drawn from the ith PSU ($i=1,2,\ldots N$). Two sub-samples of ssu's each of size m'_i were then drawn from each of the two selected samples of ssu's in the ith PSU. These were kept fixed,

that is, whenever the ith PSU is in the sample, the samples and sub-samples of ssu's are the ones drawn earlier. All possible samples of 2 PSU's were drawn from each population using the following two sampling procedures: (i) systematic selection of PSU's after arranging the units at random; (ii) selecting the first PSU using probabilities P_i (i=1,2,...N) and the second with equal probability without replacement. Values of m_i' under these procedures are shown in Tables 1 and 2.

Table 1

Values of M_i, Y_i, P_i, m_i, S^2_{wi} and m_i'

N=6 n=2

PSU	M_i	Y_i	P_i	m_i	S^2_{wi}	m_i' Sampling Procedure 1	m_i' Sampling Procedure 2
1	18	82	.1586	8	0.967	6	6
2	26	116	.2128	10	1.938	6	7
3	17	77	.1257	8	1.514	6	6
4	20	88	.1393	9	1.094	7	7
5	24	109	.2089	10	2.172	7	7
6	22	88	.1549	10	0.952	8	7

Table 2

Values of M_i, Y_i, P_i, m_i, S^2_{wi} and m_i'

N=10, n=2

PSU	M_i	Y_i	P_i	m_i	S^2_{wi}	m_i' Sampling Procedure 1	m_i' Sampling Procedure 2
1	18	82	.0932	8	0.967	7	7
2	26	116	.1250	10	1.938	8	8
3	17	77	.0732	8	1.514	7	7
4	20	88	.0818	9	1.094	8	8
5	24	109	.1227	10	2.172	8	8
6	22	88	.0909	10	0.952	8	8
7	17	76	.0784	6	1.639	5	5
8	18	76	.0909	8	1.947	7	7
9	24	111	.1204	10	1.461	8	8
10	24	112	.1227	10	1.275	8	8

The exact bias in $v_2(\hat{Y})$ was computed under each of these selection procedures for both the populations. The bias was also evaluated under a third procedure which is simple random sampling of 2 PSU's without replacement. The bias as a percentage of $V(\hat{Y})$ is shown in Table 3.

Table 3

Bias in $v_2(\hat{Y})$

Sampling Procedure	N=6, n=2			N=10, n=2		
	$V(\hat{Y})$	$Ev_2(\hat{Y})$	% Bias	$V(\hat{Y})$	$Ev_2(\hat{Y})$	% Bias
1	1418.9	2198.7	54.9	4050.2	5838.5	44.1
2	1519.3	2221.3	46.2	5315.2	6064.2	14.1
3	3675.9	5659.8	53.9	13204.6	17379.1	31.6

It is seen from the above table that the bias in $v_2(\hat{Y})$ is substantial, and further that the bias decreases as first stage sampling fraction decreases. Also, procedure 1 which is generally more efficient than procedure 2 has larger bias.

Estimates of the variance of $\hat{Y}$ were computed using each of the expressions (2.3), (2.4) and (2.8) from the samples and sub-samples drawn. For the first population, 30 values each of $v_1(\hat{Y})$ and $v_2(\hat{Y})$ and 60 values of $v_3(\hat{Y})$ were computed for each of the first two procedures. For the second population, 90 values each of $v_1(\hat{Y})$ and $v_2(\hat{Y})$ and 180 values of $v_3(\hat{Y})$ were computed for procedure 1 only. The estimated variances of $v_1(\hat{Y})$ and $v_2(\hat{Y})$ and $v_3(\hat{Y})$ are based on these values. Table 4 shows the ratio of the estimated variance of $v_1(\hat{Y})$ to the estimated variance of $v_2(\hat{Y})$ and also $v_3(\hat{Y})$ under the first two procedures for N=6 and n=2 and under procedure 1 for N=10 and n=2.

Table 4

Relative Efficiencies

Procedure	N=6, n=2 $v_1(\hat{Y})$	$v_2(\hat{Y})$	$v_3(\hat{Y})$	N=10, n=2 $v_1(\hat{Y})$	$v_2(\hat{Y})$	$v_3(\hat{Y})$
1	1.00	0.56	0.69	1.00	0.69	0.92
2	1.00	0.24	0.94			

It is seen from the above table that $v_3(\hat{Y})$ performs reasonably well.

RESUME

Même quand on fait un tirage sans remise à la première étape d'un sondage à plusieurs degrés, on calcule souvent la variance de l'estimation totale en supposant qu'on a procédé à une remise. Ce procédé introduit donc une erreur dans l'estimation, erreur qui n'est pas négligeable. Dans cet article, on présente une méthode qui permet d'obtenir des estimations sans biais de la variance en n'utilisant que les totaux estimés à l'échelle des unités primaires d'échantillonnage lorsque les échantillonnages de la deuxième étape et des suivantes sont de nature aléatoire simple sans remise. Cette technique a recours à des sous-échantillons établis à partir des unités déjà choisies du deuxième degré et des degrés suivants.

REFERENCES

[1] Des Raj, "The Use of Systematic Sampling With Probability Proportionate to Size in a Large Scale Survey", Journal of the American Statistical Association, 59 (March 1964), 251-255.

[2] Des Raj, "Sampling Theory", New York: McGraw Hill, Inc. 1968.

[3] Durbin, J., "Some Results in Sampling Theory When the Units Are Selected With Unequal Probabilities", Journal of the Royal Statistical Society, Series B, 15, (1953), 262-269.

[4] Durbin, J., "Design of Multi-Stage Surveys for the Estimation of Sampling Errors", Applied Statistics, Vol. 16, No. 2 (1967), 152-164.

[5] Harlety, H.O., and Rao, J.N.K., "Sampling With Unequal Probability and Without Replacement", Annals of Mathematical Statistics, 33 (1962), 350-374.

[6] Som, R.K., "A Manual of Sampling Techniques", London: Heinemann Educational Books Ltd., 1973.

YUKON LABOUR FORCE SURVEY - METHODOLOGY OF A FEASIBILITY STUDY

D. Royce
Household Surveys Development Division

This article summarizes the findings of a study of the feasibility of an on-going labour force survey in the Yukon Territory. The major aspects of methodology considered are the choice of a sampling frame and the determination of a sample size and allocation. It is shown that area sampling would be preferable to the use of available lists, although substantial field testing would be required because of conditions particular to the Yukon. It is also observed that sampling fractions as high as 15% may be required to produce basic labour force data, because of the small population.

1. INTRODUCTION

In the spring of 1976, Statistics Canada entered into a contract with the Department of Indian Affairs and Northern Development (DIAND) for a feasibility study with the objective of examining, in general terms, the conceptual, operational, and methodological feasibility of an ongoing labour force survey in the Yukon. A project team, consisting of members from Special Surveys Co-ordination Division, Labour Force Survey Division, Household Surveys Development Division, and the Northern Data and Liaison Division of DIAND, was established to carry out the study. The team reported to a Content Policy Group composed of senior representatives from Statistics Canada, DIAND and the Yukon Territorial Government.

The study was not restricted to the conventions of Statistics Canada's Labour Force Survey (LFS); it was explicitly understood that the project was not an extension of the LFS into the Yukon. It was also understood that other sources of labour force information, such as administrative files, were not to be considered in this study. The team presented its report [1] in August of 1976, detailing its findings and outlining a schedule

and costs for future development work. A decision on whether to proceed is now in the hands of the Federal-Territorial Economic Planning Committee.

This article is a condensation of the methodological portion of the Team's report. It should be borne in mind, however, that the conceptual problems of a labour force survey in the Yukon are equally as difficult as the methodological issues discussed below.

2. DATA REQUIREMENTS

Earlier discussions within DIAND had identified the following requirements:

- (i) "unqualified" monthly data for the Territory as a whole on the seven main labour force characteristics, i.e., population 14 years and over, labour force, employed, unemployed, not in the labour force, participation rate and unemployment rate;
- (ii) "unqualified" monthly data on an ethnic basis (native and other) for the seven main labour force characteristics (for the Territory as a whole);
- (iii) "unqualified" data on the seven main labour force characteristics for each of two zones in the Territory (see map attached);
- (iv) "qualified" labour force data classified by sex and age group and other characteristics for the Territory as a whole; and
- (v) "unqualified" supplementary data on the Territorial population engaged in hunting, fishing and trapping.

The terms "unqualified" and "qualified" are based on the release policy of the Labour Force Survey Division. The term "unqualified" signifies that the coefficient of variation of the estimate is 16.5% or less. The term "qualified" signifies that the c.v. is between 16.6% and 33%.

3. DEVELOPMENT OF METHODOLOGY

Input to the feasibility study was obtained from survey methodologists directly responsible for the design of the Canadian Labour Force Survey, from correspondence with the U.S. Bureau of the Census about experience with the Current Population Survey in Alaska, from discussions with

senior officials of DIAND and the Yukon Territorial Government, and from a field trip to the Yukon by the project team's survey methodologist.

The most important aspects of the methodology were the choice of a frame, stratification methods and the size and allocation of the sample. Section 4 gives an evaluation of the available frames and the team's recommendation. Section 5 discusses stratification and sample allocation in light of the data requirements and the costs. Formulae are developed which are used to estimate the sampling fractions necessary to meet the objectives. Section 6 summarizes the other aspects of methodology which were considered.

For the purpose of the following discussion, it was assumed that the survey would be periodic (for example, monthly or quarterly) and that the field procedures would be similar to those of Statistics Canada's Labour Force Survey, i.e. personal visits to households by interviewers, with telephone interviewing where possible after the first visit. For more details, see [2]. Some form of sample rotation was also assumed to be a necessary part of the design.

4. SAMPLE FRAME ALTERNATIVES

The Yukon Territorial Government had suggested using their Yukon Health Care Insurance Plan (YHCIP) file. Another obvious option was an area sample. A detailed evaluation of these two was carried out during the Feasibility Study.

4.1 YHCIP File

The YHCIP file is the only comprehensive list of persons or households currently available. The existing housing records of the Yukon Territorial Government do not contain lists of dwellings or even adequate counts of dwellings. The unorganized nature of many of the settlements means that

administrative records of dwellings (such as building permits) are inadequate as a frame.

The YHCIP file is a list of all residents of the Yukon Territory who are required, by law, to register with the Health Care Insurance Plan. The file is updated on a monthly basis, and contains the resident's name, address, Social Insurance Number, sex, dependent code, date of birth, employer code and other information relevant to the plan. It is also possible to identify registered natives from information on the file.

The advantages of the YHCIP file as a frame are:

(1) The auxiliary information on age, sex, dependent status, and native status could be useful for stratification, and

(2) The use of this list would avoid the time and expense of field listing inherent in area sampling.

There are also several disadvantages of the YHCIP file as a sample frame, mostly arising from the fact that the basic unit (record) on the file is a person rather than a dwelling.

(1) If persons rather than dwellings were selected, then it becomes necessary to trace persons, even when they move, in order to maintain the sample. Since mobility is fairly common, this would likely be a serious problem. Tracing of persons is a time-consuming and expensive procedure, and would likely make production of monthly data impossible.

(2) Selection of persons rather than dwellings also means that usually only one person per household would be selected. This may mean that more households would have to be visited unless there were a strong positive correlation between members in the same household; that is, persons in the same household tending to have similar characteristics. Usually, the intra-household correlation is negative, and therefore, it is more efficient to take all members of the household.

(3) The respondent burden within a household is potentially very high because rotation of the sample could easily result in changing from one member of a household to another in the same household.

The problems associated with selecting persons make it desirable to select households instead. Selection of households could be done by using the YHCIP file to select a person, and then collecting data for everyone living in the same dwelling. This amounts to selection of households with probability proportional to household size. However, there are also problems with this approach:

(1) Most of the auxiliary information on the file would be difficult or impossible to use, since it pertains to persons rather than households.
(2) Attempting to set up a rotation procedure for households based on a list of persons would result in severe practical problems, because the household compositions are not known in advance. For example, selection of six samples of persons (for the six rotation groups) could result in a household being in two rotation groups if two persons in the same household were selected. Rotation of persons could also result in returning to the same household, a situation which would cause a very high respondent burden. In practice, an on-going survey involving sample rotation of households would not be feasible using the YHCIP file.

In addition to the above conceptual problems, there were other problems with the YHCIP file which made it unsuitable as a frame:

(1) Usable addresses exist for only about 70% of Whitehorse and hardly at all for other areas of the Yukon. Although local knowledge of the whereabouts of the inhabitants might be sufficient in some cases, some tracing of persons would be needed, which would likely delay data production.
(2) The procedures for updating the file are currently inadequate for a sampling frame for an on-going survey. The most problematic are people who move out of the Yukon; there can be as much as a year's delay in recording such moves. This can be handled by

domain estimation, but the efficiency of the sample is reduced and tracing of these people would be expensive. There is also no systematic procedure for updating changes of addresses.

In summary, use of the YHCIP file as a frame was not recommended for the following reasons:
(1) The structure was unsuitable for an on-going household survey.
(2) The file was not sufficiently complete and up to date.

4.2 Area Sample

Area sampling is the method used for many household surveys, including the LFS in Canada and the Current Population Survey (CPS) in the United States. The main advantages are:

(1) The physical dwelling unit, which is the final sampling unit, is relatively stable over time, thus eliminating problems of tracing and sample rotation.
(2) The coverage is relatively accurate, both of dwellings within geographical areas and of persons within dwellings.
(3) For labour force characteristics, the intra-household correlation tends to be negative, making the household an efficient cluster for sampling.
(4) The approach is amenable to multi-stage sampling, which is often necessary to reduce field costs.

An additional advantage in the Yukon context would be that of comparability with the rest of the country.

The major disadvantage of the area sampling approach is the time and expense involved in field listing. In the Yukon context, it was also expected that there might be technical problems with this step.

A field trip to the Yukon by the team's methodologist in early May of 1976 indicated that designing an area sample for the city of Whitehorse would likely present few problems. Dwellings are usually quite distinct, and

most have a street name and number. The only foreseeable problem might be in the large land area covered by the city, thus increasing travel costs.

More difficult problems would likely be found in the smaller settlements. In some cases, dwellings cannot always be immediately and uniquely identified. During one personal visit, it was quite hard at times to distinguish between living quarters and a shack for tools and other equipment. Previous experience in the Yukon with the Family Expenditure pilot had also shown a distressing tendency for some of the dwelling units to move from time to time. Dwellings appear scattered rather than built according to a pre-defined plan, consequently very precise descriptive addresses would be required.

Outside of the small settlements, the problems become even more severe. Dwellings along the side of the road are quite often hidden in the bush and accessible only by a long lane-way. According to DIAND, fairly extensive personal knowledge of these dwellings is available in the Yukon Government but a great deal of checking and testing would still be necessary. There also exist temporary hunting and fishing camps which are used only during the summer months. Special procedures for handling these, for instance interviewing once a year, and assuming vacant during the winter, would be required.

Another problem which the Alaska CPS encountered was the lack of natural boundaries which could be used for delineating segments to be sampled. The solution was to define segments in terms of a village and the land area around that village.

Finally, coverage of persons within dwellings is likely to be a problem in the rural areas. Hunters, fishermen and trappers are often absent from home for extended periods of time because of their work. Some method for handling these people would be necessary.

Some field testing would definitely be necessary to attempt to solve these problems. However, based on experience in the rest of the counting and the CPS experience in Alaska, the problems of area sampling seem to be much less severe than those of the YHCIP file. It was recommended that future work concentrate solely on the area sampling approach.

5. STRATIFICATION AND SAMPLE ALLOCATION

The Yukon Government has defined two Zones in the Territory, Zone A consisting of Whitehorse, Haines Junction, Dawson, Mayo, Faro, and Watson Lake and Zone B consisting of the remainder of the Territory. Since separate data for Zone A and Zone B were desired, this would be the obvious first level of stratification. Zone A would be further stratified into Whitehorse and one or more strata containing the other Zone A communities. Zone B would likely be formed of non-self representing units and would require at least three stages of sampling (primary sampling units, segments and households) in order to reduce travel costs. Stratification of the primary sampling units would be done to some extent, but because of the small population (about 1,400 households) very few strata could be formed.

The allocation of the sample to the strata depends on both the variance in each stratum and the enumeration costs in each stratum. Mathematically, the sample size and allocation is determined by minimizing the total cost subject to certain constraints on the variances of the sample estimates. In this instance, we will minimize the total cost subject to the condition that unqualified data for the labour force characteristics can be obtained for the Yukon Territory as a whole for Zone A and Zone B separately, for the Native and Non-Native populations separately, and that the number of persons employed in Hunting, Fishing and Trapping can be measured at an unqualified level.

We consider a stratified sampling scheme where units are selected in clusters. (This is the type of scheme being proposed for the Yukon situation). The estimate of the total number of units in the population (X) having the characteristic to be estimated is:

$$\hat{X} = \sum_{h=1}^{L} \hat{X}_h = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} x_{hi}$$

where L is the number of strata

N_h is the size of stratum h

n_h is the sample size in stratum h and

x_{hi} = 1 if unit i of stratum h has the characteristic

0 otherwise.

The variance of X is given by

$$V(\hat{X}) = \sum_{h=1}^{L} V(\hat{X}_h) = \sum_{h=1}^{L} F_h N_h^2 \frac{N_h - n_h}{N_h - 1} \frac{P_h (1 - P_h)}{n_h}$$

where F_h, the design effect, is a factor multiplying the Simple Random Sampling Without Replacement variance to allow for the clustering effect and any further sub-stratification, and P_h is the proportion of units in stratum h having the characteristic.

The mathematical constraints implied by the "unqualified data" conditions are:

(1) Coefficient of Variation of Estimate $\frac{\sqrt{V(\hat{X})}}{X} \leq 0.165$

for all estimates $\hat{X}$ which are to be unqualified for the Territory as a whole.

(2) Coefficient of Variation of Estimate $\sqrt{\frac{V(\hat{X}_h)}{X_h}} \leq 0.165$

for all estimates which are to be unqualified for stratum h. (In this case, the two strata are Zone A and Zone B). Letting c_h be the cost of obtaining an interview in stratum h, the problem can be formulated as:

$$\text{Minimize } C = \sum_{h=1}^{L} c_h n_h$$

subject to:

1) $V(\hat{X}) = \sum_{h=1}^{L} F_h N_h^2 \left(\frac{N_h - n_h}{N_h - 1}\right) \frac{P_h(1-P_h)}{n_h} \leq (0.165)^2 \left(\sum_{h=1}^{L} P_h N_h\right)^2$

2) $V(\hat{X}_h) = F_h N_h^2 \left(\frac{N_h - n_h}{N_h - 1}\right) \frac{P_h(1-P_h)}{n_h} \leq (0.165)^2 P_h^2 N_h^2$

3) $1 \leq n_h \leq N_h$

where constraint 1) applies to all characteristics to be published at the Territory level (e.g. Unemployed, Unemployed Natives, Engaged in Hunting, Fishing and Trapping), constraint 2) applies to all characteristics to be published at the stratum level (e.g. unemployed in Zone A and in Zone B), and constraint 3) expresses the restriction that the sample size in each stratum cannot be zero and also cannot be larger than the stratum size itself.

Letting $r_h = \frac{1}{n_h} - \frac{1}{N_h}$, the above can be reformulated as:

$$\text{Minimize} \quad C = \sum_{h=1}^{L} \frac{N_h \; c_h}{1 + N_h \; r_h}$$

Subject to:

1) $\Sigma \; (F_h \; N_h^2 \; S_h^2) \; r_h \leq (0.165)^2 \; (\quad P_h \; N_h)^2$

2) $(F_h \; S_h^2) \; r_h \qquad \leq (0.165)^2 \quad P_h^2 \qquad h = 1,2$

3) $0 \leq r_h \leq 1 - \frac{1}{N_h}, h = 1,2$

where $S_h^2 = \frac{N_h}{N_h - 1} \; P_h \; (1 - P_h) \doteq P_h \; (1 - P_h)$.

This is a convex programming problem since the objective function C is a separable convex function of the r_h and the constraints are linear in r_h. Therefore, given values of c_h, F_h, N_h and P_h (for various characteristics), the necessary sample sizes to achieve the desired reliability at minimum cost can be determined.

Values of c_h are not required for working with the constraints, which is our main interest at this point. Values of F_h can be estimated from experience with the LFS in other parts of Canada. The values of N_1 and N_2 were obtained from the 1971 Census of Population. The most critical characteristic for constraint (1) is the Native Unemployed.

For constraint (2), it is the Unemployed (broken down by Zone A/Zone B but not Native/Non-Native). Using a value of 10% Native Unemployed, it was found that sampling ratios of about one in six in Zone A and one in two in Zone B would be required. This would also guarantee unqualified estimation of Unemployment down to a true proportion of 4.0% in each of Zone A and Zone B.

Sampling ratios based on breakdowns other than Native/Non-Native and Zone A/Zone B were calculated using the binomial variance formula with a design effect of 2.0. It was not possible to use the more exact formulae used above, since appropriate values of P_1 and P_2 were unknown in most cases. The results are shown in Table 1.

Table 1: Sampling Fractions for Various Data Requirements

Data Requirements	Sampling Fraction Estimated For	
	CV = 16.5%	CV = 33%
	(Percentage of Population)	
(i) Seven main labour force characteristics for the Territory	13%	n.a.
(ii) As in (i) for native and non-native	45%	17%
(iii) As in (i) for each of two zones	16% in Zone A 42% in Zone B	5% in Zone A 15% in Zone B
(iv) As in (i)		
- by sex	24%	7%
- by age groups	63%	30%
- by marital status	77%	43%
- by educational attainment	77%	32%
- by industry	100%	100%
- by occupation	83%	59%
- by class of worker	91%	77%
- by educational activity	56%	23%
(v) Hunting, Fishing and Trapping data for entire Territory	100%	100%

Examination of the tables where the CV is 16.5% shows that with a sampling fraction of 15%, one could publish an estimate of the Uemployed only for the population as a whole. No other breakdowns could be published. However, some of the data could be available for analysis, as shown by the set of tables where the CV is 33%. For a sampling ratio of 15%, Unemployed could be measured by sex, by collapsed age groups, by zone, by marital status (excluding "Other"), by collapsed education attainment (combining "Some University" and "University Degree"). It could not be measured by detailed age groups, or by educational activity or by ethnicity, or by industry, or by occupation or by class of worker.

The above discussion has mentioned only sampling error, since it is the type of error that can be most easily treated mathematically. However, in addition there will be present some amount of non-sampling error, due to such sources as non-response, training or attitudes of the interviewers, faults in the questionnaire design, processing errors and so on. It should be kept in mind when considering sample sizes and the corresponding amount of error that the actual amount of error in the estimate may be more than that due to sampling error alone.

Summary of Sample Size Evaluation

1) The initial data requirements cannot be met because they imply too high a sampling ratio.
2) A sampling ratio of about 15% would allow a publishable estimate of the number of unemployed for the Yukon as a whole. It would also allow for breakdowns by sex, major age groups, zone, marital status and major educational attainment at a coefficient of variation between 16.5% and 33%.
3) In addition to the sampling error, there will be some degree of non-sampling error which will tend to make the estimates less reliable than indicated by the sampling error alone.

6. OTHER CONSIDERATIONS

The Feasibility Study also examined two other aspects of methodology, that of appropriate stages of sampling in each of Zone A and Zone B, and that of estimation and variance estimation. In both cases, it was concluded that the design would be very similar to the LFS itself. The only exception was the possible use of the YHCIP file as a source of external estimates for improving the survey estimates.

7. ACKNOWLEDGEMENT

The author wishes to acknowledge the contribution of Project Team members, C. Scott (SSCD), M. Webber (LFSD), and R. Raghunathan (DIAND) in the preparation and writing of the report on which this paper is based. Acknowledgement is also made to R. Ryan (SSCD) for contributions during the early work of the Project Team.

RESUME

Cet article résume les résultats d'une étude sur la faisabilité d'une enquête permanente sur la population active dans le Territoire du Yukon. Les grandes lignes de la méthodologie envisagée sont le choix d'un plan de sondage et la détermination de la taille de l'échantillon et de sa répartition. On a démontré qu'il serait préférable d'utiliser une méthode aréolaire de sondage plutôt que d'avoir recours aux listes disponibles, malgré la nécessité de procéder à un important essai sur le terrain à cause des conditions de travail particulières au Yukon. On a également montré qu'à cause de la faible population de cette région, on aurait peut-être besoin de fractions de sondage allant jusqu'à 15% afin d'obtenir les données de base sur la population active.

REFERENCES

[1] "A Labour Force Survey in the Yukon Territory - Feasibility Report", Statistics Canada, August, 1976.

[2] Platek, R. and Singh, M.P., "Methodology of Canadian Labour Force Survey", Statistics Canada, 1976.

c. 3

DATE DUE

SURVEY METHODOLOGY/TEC

June/juin 1976 Vol. 2

A Journal produced by Statistical Ser

Publié par le secteur des services st

CONTEN

SURVEY METHODOLOGY/TECHNIQUES D'ENQUÊTE

December/decembre 1976 Vol. 2 No. 2

A Journal produced by Statistical Services Field, Statistics Canada.

Publié par le secteur des services statistiques, Statistique Canada.

C O N T E N T S